TOWARDS AUTOMATING SLEEP STAGE SCORING TO DIAGNOSE SLEEP DISORDERS

by
Kristin Maria Gunnarsdottir

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland

May, 2016

Abstract

Overnight polysomnography (PSG) is an important tool used to characterize sleep and the gold standard procedure for diagnosing many sleep disorders. PSG is a non-invasive procedure that collects various physiological data, such as EEG, EMG, EOG and ECG signals. The data is then scored in a subjective, laborious and time-consuming process by sleep specialists who assign a sleep stage to every 30-second window of the data according to predefined scoring rules by the American Academy of Sleep Medicine (AASM). Finally, clinicians make a diagnosis based on this annotated data. Consequently, the current process is heavily dependent upon human factors, which can result in poor agreement between expert scorers, but inter-scorer reliability has been found to be only around 82%.

In this study we developed an automatic sleep stage scoring method, using a likelihood ratio decision tree classifier, with the goal of improving the speed, reliability, accuracy and cost efficiency of the current PSG scoring process. The algorithm was developed using the AASM Manual for Scoring Sleep. We extracted features from various physiological recordings of the PSG, based on the predefined rules of the AASM Manual. The features were computed for each 30-second epoch, in either the time or the frequency domain. The most useful features were selected by looking at probability distributions for each metric conditioned on the sleep stage, and identifying the features giving the greatest separation between stages. Examples of meaningful features include the power in different frequency bands of EEG signals, EMG energy per epoch, and number of spindles per epoch, to mention a few. These features were then used as inputs to the classifier which assigned each epoch one of five possible stages: N3, N2, N1, REM or Wake.

The automatic scoring was trained and tested on PSG data from 39 healthy individuals (age range: 24.2 ± 3.1 years) with no sleep disturbances. The overall scoring accuracy was 76.97% on the test set. Some of the stages, such as stage N2, have more distinctive characteristics and thus yielded a higher per-stage scoring accuracy, whereas the other stages, for example stages N1 and REM, got confused more easily, resulting in lower per-stage accuracies. As expected, most misclassifications occurred between adjacent sleep stages. Although this accuracy may at first seem low, it is likely that the stages that the tool classified inaccurately may be sleep stages that contribute to inter-scorer reliability. Therefore, we see this tool as assisting sleep scorers to enhance efficiency with the further goal of eventually improving inter-scorer reliability.

Sleep stage scoring provides an important basis for diagnosis of sleep disorders in general. However, the detection of sleep disturbances is very costly and time-consuming, and relies on subjective measures. Automating the scoring process improves the efficiency and consistency of scoring procedures and offers a way to diagnose sleeping disorders in a more robust, quantitative manner.

Thesis Committee

Dr. Sridevi V. Sarma (advisor)

Assistant Professor, Department of Biomedical Engineering

Johns Hopkins University

Dr. Charlene E. Gamaldo

Associate Professor, Neurology

Johns Hopkins Medicine

Dr. Rachel Marie. E. Salas

Associate Professor, Neurology

Johns Hopkins Medicine

Acknowledgements

First of all, I would like to express my sincere gratitude to my advisor, Dr. Sri Sarma for the continuous support, guidance and motivation throughout my Master's studies. Her enthusiasm and positive encouragement have influenced my desire to pursue PhD studies in Biomedical Engineering and I truly could not have imagined having a better advisor and mentor.

Besides my advisor, I would like to thank my other mentors, Dr. Charlene Gamaldo and Dr. Rachel Salas for contributing their time and expertise throughout the project. Their insightful comments and feedback, and always positive attitude, were critical to the success of this project. I am grateful to have had the opportunity to work with such an amazing team of mentors.

I would also like to acknowledge the faculty involved in the Biomedical Engineering Master's Program at Johns Hopkins. Special thanks to Samuel Bourne for all his suggestions and help during the past two years.

I am extremely thankful for all the new friendships I have made during my Master's studies. I want to thank all my friends in Baltimore for sharing this journey with me, through the good times and the tough times, and for all the fun we had together. The past two years would not have been the same without them.

Most importantly I want to thank my parents and my sisters for always believing in me and for their endless love and support. Without their patience, understanding and encouragement I never would have made it through.

Table of Contents

Abstract ii

Thesis Committee iv

Acknowledgements v

List of Tables viii

List of Figures viii

1 Introduction 1

1.1 Overnight Polysomnography 1

1.2 Sleep Stages 2

1.2.1 Stage Wake 3

1.2.2 Stage N1 4

1.2.3 Stage N2 5

1.2.4 Stage N3 6

1.2.5 Stage REM 6

1.3 Sleep Stage Scoring and Hypnograms 7

1.4 Automating Sleep Stage Scoring 9

2 Methods 15

2.1 Database 15

2.2 Data Acquisition 16

2.3 Data Analysis 16

2.3.1 Human Expert Scoring 16
2.3.2 Automated Sleep Stage Scoring Algorithm 17
2.3.2.1 Data Preprocessing 18
2.3.2.2 Feature Extraction 19
2.3.2.3 Threshold Determination 26
2.3.2.4 Automatic Sleep Stage Classification 27
3 Results 30
3.1 Quantitative Feature Distributions 30
3.2 Training Set Results 39
3.3 Test Set Results 41
4 Discussion 45
5 Conclusions and Future Work 52
5.1 Conclusions 52
5.2 Future Work 53
6 References 56
6.1 Appendices 56
Appendix A : Test Set Hypnograms 56
Appendix B : Test Set Run-Times 66
6.2 Bibliography 67
7 Curriculum Vitae 73

List of Tables

Table 1. Sleep frequency bands as defined by the AASM Scoring Manual. 2

Table 2. Population statistics. 15

Table 3. Filter settings of the PSG signals. 18

Table 4. The features used in the classifier. 25

Table 5. Scoring results of the training set. 40

Table 6. Scoring results of the test set. 42

Table 7. Sleep vs. Wake scoring results of the test set. 44

Table A-1. The scoring run-times of all test subjects. 66

List of Figures

Figure 1. Typical activity of the EOG, EEG and EMG recordings during Wake. 3

Figure 2. Typical activity of the EOG, EEG and EMG recordings during N1. 4

Figure 3. Typical activity of the EOG, EEG and EMG recordings during N2. 5

Figure 4. Typical activity of the EOG, EEG and EMG recordings during N3. 6

Figure 5. Typical activity of the EOG, EEG and EMG recordings during REM. 7

Figure 6. An example of a hypnogram. 8

Figure 8. A block diagram of the automatic sleep stage classification procedure. 9

Figure 7. A comparison of the performance of our algorithm and a few existing scoring methods. 12

Figure 9. Electrode placements. 16

Figure 10. Cross-correlations of two EOG signals. 21

Figure 11. Autocorrelations of an EOG signal. 22

Figure 12. Probability density functions and the corresponding decision thresholds..... 26

Figure 13. A flowchart of the automatic scoring process. 27

Figure 14. Probability distributions of each sleep stage for $EOG_1^{0.3-35}$. 31

Figure 15. Probability distributions of each sleep stage for Delta Power in group 1. 31

Figure 16. Probability distributions of each sleep stage for Beta Power in group 1..... 32

Figure 17. Probability distributions of each sleep stage for Maximum EMG energy in group 1. 32

Figure 18. Probability distributions of each sleep stage for Theta Power in group 1..... 33

Figure 19. Probability distributions of each sleep stage for Alpha Power in group 1. 33

Figure 20. Probability distributions of each sleep stage for $EOG_2^{0.1-0.45}$ in group 1. 34

Figure 21. Probability distributions of each sleep stage for Maximum Spindle Duration in group 1. 34

Figure 22. Probability distributions of each sleep stage for Number of Spindles in group 1. 35

Figure 23. Probability distributions of each sleep stage for Delta Power in group 1. 35

Figure 24. Probability distributions of each sleep stage for $EOG_3^{0.3-0.45}$ in group 1..... 36

Figure 25. Probability distributions of each sleep stage for Maximum EMG energy in group 2. 36

Figure 27. Probability distributions of each sleep stage for Alpha Power in group 2. 37

Figure 26. Probability distributions of each sleep stage for Theta Power in group 2..... 37

Figure 28. Probability distributions of each sleep stage for $EOG_2^{0.1-0.45}$ in group 2..... 38

Figure 29. Confusion matrix using the training data set. 40

Figure 30. Confusion matrix using the test data set. 42
Figure 31. Sleep vs. Wake confusion matrix using the test data set. 44
Figure A-1. Hypnograms and scoring accuracy for test subject 1. 56
Figure A-2. Hypnograms and scoring accuracy for test subject 2. 57
Figure A-3. Hypnograms and scoring accuracy for test subject 3. 57
Figure A-4. Hypnograms and scoring accuracy for test subject 4. 58
Figure A-5. Hypnograms and scoring accuracy for test subject 5. 58
Figure A-6. Hypnograms and scoring accuracy for test subject 6. 59
Figure A-7. Hypnograms and scoring accuracy for test subject 7. 59
Figure A-8. Hypnograms and scoring accuracy for test subject 8. 60
Figure A-9. Hypnograms and scoring accuracy for test subject 9. 60
Figure A-10. Hypnograms and scoring accuracy for test subject 10. 61
Figure A-11. Hypnograms and scoring accuracy for test subject 11. 61
Figure A-12. Hypnograms and scoring accuracy for test subject 12. 62
Figure A-13. Hypnograms and scoring accuracy for test subject 13. 62
Figure A-14. Hypnograms and scoring accuracy for test subject 14. 63
Figure A-15. Hypnograms and scoring accuracy for test subject 15. 63
Figure A-16. Hypnograms and scoring accuracy for test subject 16. 64
Figure A-17. Hypnograms and scoring accuracy for test subject 17. 64
Figure A-18. Hypnograms and scoring accuracy for test subject 18. 65
Figure A-19. Hypnograms and scoring accuracy for test subject 19. 65

1 Introduction

Sleep is a basic human need. It is an important aspect of health and wellbeing and is strongly related to overall quality of life [1]. According to the Institute of Medicine, it is estimated that 50 to 70 million Americans suffer from a chronic sleep disorder adversely affecting daily functioning and overall health [2]. Today, sleep disorders are generally diagnosed through clinician interviews or questionnaires such as the Pittsburgh Sleep Quality Index, which are based on subjective assessments and are as such prone to inaccurate diagnostics.

1.1 Overnight Polysomnography

Overnight polysomnography (PSG) is the gold standard to diagnose sleep apnea and REM sleep behavior disorder, but can offer important support for other clinical diagnosis such as restless legs syndrome, parasomnias and sleep related movement disorders. PSG is usually conducted in a hospital or a sleep center and involves collecting multiple physiological recordings in a non-invasive way. This includes electroencephalography (EEG), electromyography (EMG), electrooculography (EOG) and electrocardiography (ECG) as well as other signals such as nasal airflow, respiratory effort and body temperature. The data is then manually scored in a subjective and time-consuming process by sleep specialists. This process is quite laborious because it primarily entails sleep stage assignment in 30-second windows (epochs) using guidelines established by the American Academy of Sleep Medicine (AASM). A seasoned registered sleep technologist at the

Johns Hopkins Sleep Center takes for instance around 30 minutes to 1.5 hours on average to stage a full night sleep study across all physiologic channels monitored [Gamaldo, C.E. & Salas, R.M.E., personal communication, April 28, 2016]. Finally, clinicians review the scored study in order to evaluate the evidence for various sleep disorders. As a result, the current process is heavily dependent upon human factors that is fraught with variable and often poor inter-scorer reliability.

1.2 Sleep Stages

The two main types of sleep are rapid eye movement (REM) and non-rapid eye movement (non-REM) sleep. Throughout the night, non-REM and REM sleep alternate in a cyclical fashion with an average duration of one cycle typically around 45-90 minutes [3]. According to the AASM Manual for the Scoring of Sleep and Associated Events the sleep cycle of adults consists of five stages: Wakefulness, REM stage and three non-REM stages: N1, N2 and N3 [4]. Each sleep stage is characterized by the presence (or the absence) of certain wave forms and events, most commonly observed in the EEG signals. For sleep analysis, the EEG activity is typically separated into four characteristic frequency bands: delta, theta, alpha and beta (Table 1). As non-REM sleep progresses, the brain becomes less responsive to external stimuli and brain waves become slower and more synchronized causing the brain to increase its arousal threshold, thus making it harder to wake an individual up from

Table 1. Sleep frequency bands as defined by the AASM Scoring Manual.

Frequency band	Frequency range (Hz)
Delta	0-4
Theta	4-7
Alpha	8-13
Beta	13-30

sleep. Most slow wave non-REM sleep occurs in the first third of the night whereas REM sleep episodes predominate the last third of the night. Below is a description of the main characteristics of each sleep stage as defined by the AASM Manual.

1.2.1 Stage Wake

Stage Wake represents the waking state and ranges from full alertness through early stages of drowsiness. During wakefulness, the majority of individuals with eyes closed will demonstrate alpha rhythm, and consequently Wake is scored when alpha rhythm is present over the occipital region for more than 50% of the epoch. The EOG signals may demonstrate rapid eye blinks, with a synchronized positive polarity across both eyes, at a frequency of around 0.5-2 Hz and while rapid eye movements are characteristic of REM, sleep they can also occur when subjects have their eyes open in stage Wake. As drowsiness develops, slow eye movements may occur. According to the AASM Manual, an epoch with major body movements should be scored as Wake if alpha rhythm is present or if an scorable Wake epoch either precedes or follows the epoch [4]. The chin EMG during Wake can be of variable amplitude but is usually higher than during sleep. Typical activity of the EOG, EEG and EMG signals in a single Wake epoch can be seen in Figure 1.

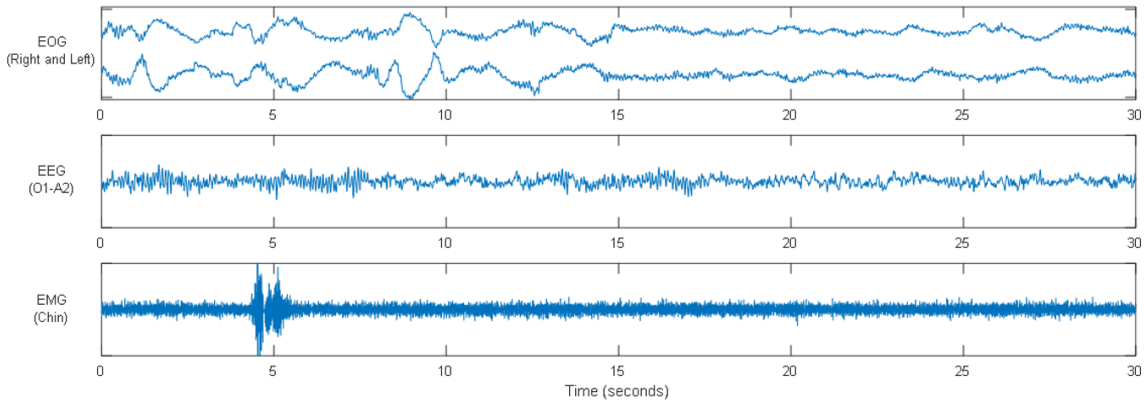

Figure 1. Typical activity of the EOG recordings (top), the EEG recordings (middle) and the EMG recordings (bottom) during Wake.

1.2.2 Stage N1

Stage N1 embodies the transition from Wake to sleep. N1 marks sleep onset, defined as the first epoch scored as any stage other than stage Wake. Stage N1 is also scored when alpha rhythm is attenuated and replaced by low amplitude, mixed frequency activity for more than 50% of the epoch. The chin EMG amplitude is variable, but often lower than in stage Wake. Additionally, sharply contoured waves (vertex sharp waves) with duration of less than 0.5 seconds may be present in the EEG and the EOG signals will often show slow eye movements, defined as conjugate, sinusoidal eye movements with an initial deflection usually lasting for more than 500 msec. However, neither vertex sharp waves nor slow eye movements are required for scoring stage N1 [4]. Figure 2 demonstrates an example of the activity observed in the EOG, EEG and EMG signals during N1.

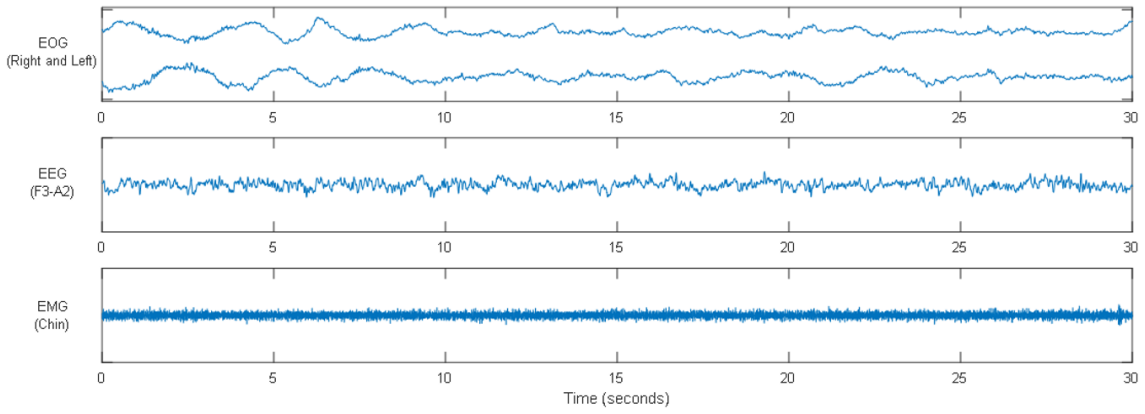

Figure 2. Typical activity of the EOG recordings (top), the EEG recordings (middle) and the EMG recordings (bottom) during N1.

1.2.3 Stage N2

The two main characteristics of stage N2 are K complexes and sleep spindles (Figure 3). A K complex is defined as a negative sharp wave immediately followed by a positive component standing out from the background EEG, with a total duration longer than 0.5 seconds. K complexes are usually maximal in amplitude over the frontal regions. Sleep spindles, on the other hand, are trains of distinct waves with frequency 11-16 Hz (most commonly 12-14 Hz) with a duration of more than 0.5 seconds and a maximal amplitude in the central derivations. Epochs with low-amplitude, mixed frequency EEG activity, without any K complexes or sleep spindles, should also be scored as N2 if they are preceded by epochs containing K complexes or sleep spindles. In stage N2 the chin EMG is of variable amplitude, but usually lower than in Wake and may be as low as during REM sleep [4].

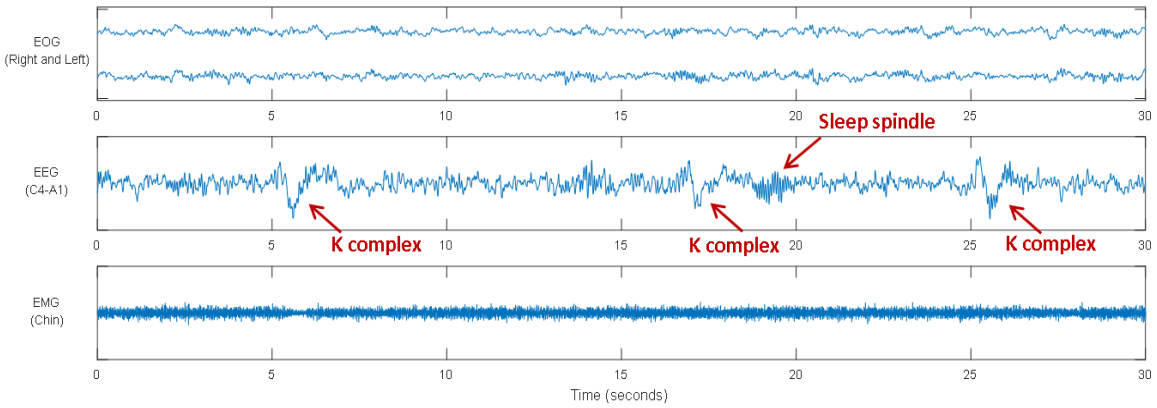

Figure 3. Typical activity of the EOG recordings (top), the EEG recordings (middle) and the EMG recordings (bottom) during N2.

1.2.4 Stage N3

Stage N3 is often referred to as slow wave or deep sleep and is characterized by slow wave, delta activity, of frequency 0.5-2 Hz and peak-to-peak amplitude greater than 75μV taking up more than 20% (6 seconds) of the epoch. Eye movements are not typically seen during N3 sleep, and although variable, the chin EMG amplitude is often lower than in N2 sleep and sometimes as low as during REM sleep [4]. The arousal threshold is often the highest during this stage of sleep, hence the common reference to deep sleep. Typical activity of the EOG, EEG and EMG signals during N3 can be seen in Figure 4.

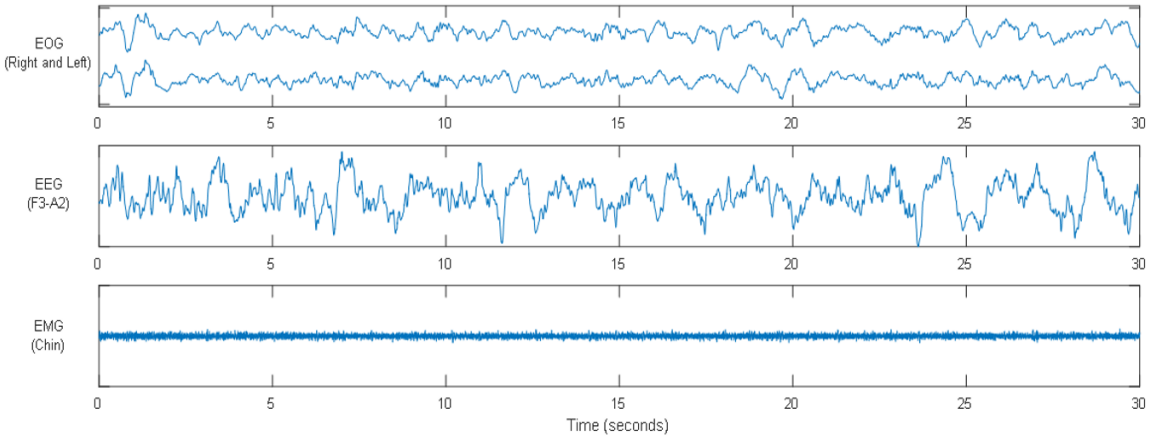

Figure 4. Typical activity of the EOG recordings (top), the EEG recordings (middle) and the EMG recordings (bottom) during N3.

1.2.5 Stage REM

REM sleep typically comprises about 20-25 % of total sleep in healthy adults. The REM stage is characterized by rapid eye movements (REM), defined in the AASM Manual as conjugate, irregular, sharply peaked eye movements with an initial deflection usually lasting for less than 500 msec and opposite polarity between eyes [4]. The EEG signal consists of low-amplitude, mixed frequency activity with trains of sharply contoured or

triangular waves, called sawtooth waves, commonly present. The chin EMG tone is low, with the baseline EMG activity no higher than in any other sleep stage and usually at the lowest level of the entire recording. Figure 5 demonstrates a typical example of the activity seen in the EOG, EEG and EMG signals during REM sleep.

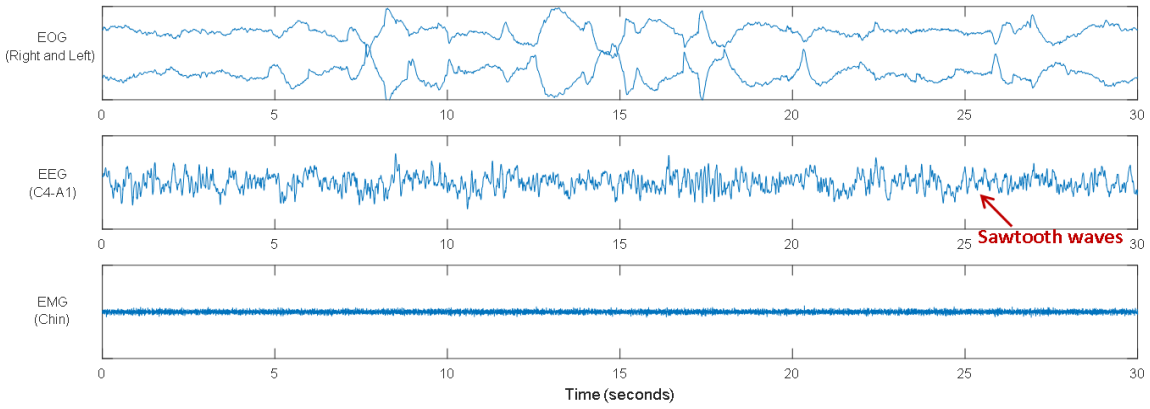

Figure 5. Typical activity of the EOG recordings (top), the EEG recordings (middle) and the EMG recordings (bottom) during REM.

1.3 Sleep Stage Scoring and Hypnograms

The successive visual scoring, by 30-second epochs, of whole night recordings leads to the representation of the temporal distribution of the five sleep stages called a hypnogram (Figure 6). A hypnogram reveals the cyclical pattern of sleep as it shifts between the different stages of sleep and wakefulness. During the night, a normal sleeper moves between the different sleep stages in a relatively predictable pattern, transitioning between REM and the sequential stages of non-REM sleep. Consequently, polysomnography provides a lot of information about an individual's sleep and is important for assessing sleep quality and a variety of sleep disturbances.

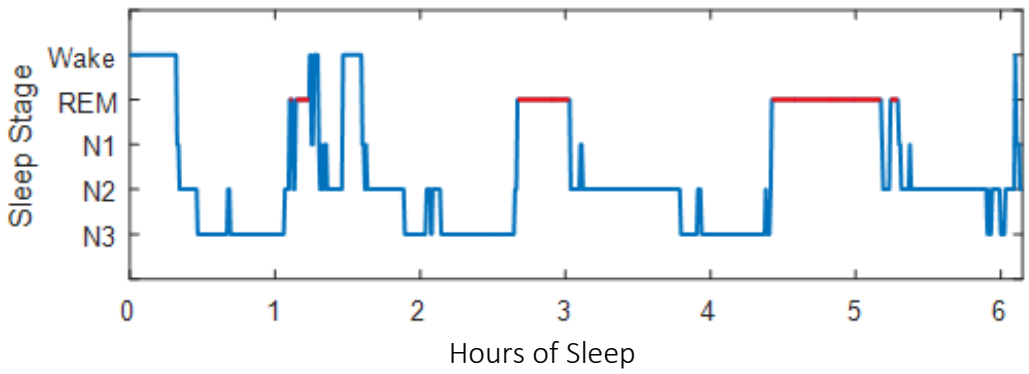

Figure 6. An example of a hypnogram.

At present sleep stage annotation is very costly, time-consuming and subjective and requires a great amount of effort from a trained specialist. An experienced sleep technologist takes around 30 minutes to 1.5 hours on average to stage a full night sleep study and the subjective nature of the annotation process can give rise to scoring errors, as even well trained specialists can be uncertain about the presentation of certain states. Moreover, studies have shown that, even among experts, the inter-scorer reliability is only about 82% [5]. As stated by Silber et al. [6] "No visual based scoring system will ever be perfect, as all methods are limited by the physiology of the human eye and visual cortex, individual differences in scoring experience, and the ability to detect events viewed using a 30-secon epoch". Thus, a more efficient and rigorous method that facilitates the scoring procedure and provides a consistent and accurate scoring of sleep stages is needed.

1.4 Automating Sleep Stage Scoring

While visual scoring still remains the gold standard of sleep stage scoring, numerous research studies have been conducted to automate the scoring process in order to improve the efficiency and consistency of scoring procedures and offer a way to diagnose sleeping disorders in a more robust, quantitative manner. Automating the sleep stage scoring process has been of increasing interest to researchers and clinicians in the field of sleep medicine since 1970. Most existing scoring methods consist of five general steps: (1) PSG data collection, (2) human expert sleep stage scoring of PSG data, (3) feature extraction from the PSG recordings, (4) automatic sleep stage classification and (5) comparison of human expert and automatic sleep stage scoring (Figure 7). The different methods found in the literature vary by the PSG recordings used as well as the implementation of the feature extraction and the classification algorithm steps. Some approaches include an additional step of feature selection before applying the sleep stage classifier.

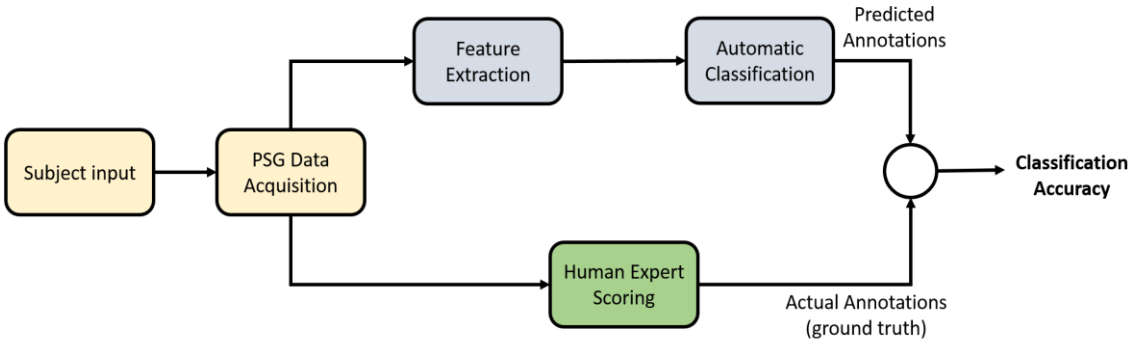

Figure 7. A block diagram of the automatic sleep stage classification procedure.

Existing scoring methods can be broadly divided into two categories; methods based on rule-based reasoning and machine learning methods based on numerical classification. The rule-based approaches [7–10] use supervised, expert-based features and classification rules derived from either the R&K manual or the more recent sleep scoring manual by the AASM. Conversely, the numerical classification methods [11–29] often make use of unsupervised, blind spectral features that do not require any prior knowledge about sleep data. Nonetheless, there exist a few numerical classifiers that also employ expert-based features or a combination of both [30–33]. Furthermore, some hybrid systems have been proposed in order to exploit the advantages of both classification methods [34–36].

Amongst the polysomnography recordings, EEG signals are by far the signals used most often [9, 11, 12, 16–18, 20–23, 26, 31–35] since they best represent the brain's activity during sleep, with different characteristics present in different sleep stages [4]. However, the EEG activity of some stages, particularly stages N1 and REM, is highly similar [37] leading to a poor classification performance for discriminating between those stages. Thus the EEG spectral information is not sufficient to distinguish between those stages, and it has been shown that other physiological signals of the PSG, such as EMG or EOG, provide important additional information than can improve the overall scoring accuracy [14]. While several studies have incorporated information extracted from the EMG in addition to the EEG signals [10, 38], others have used both the EMG and EOG recordings [7, 24, 29], which is in agreement to the AASM rules followed by human experts. Although less common, a few studies have also obtained promising results using features extracted from signals such as the EOG signals only [9, 25, 28] or the ECG and respiratory recordings [27, 30] .

Across existing methods, a variety of features have been extracted from the PSG signals including time-domain, frequency-domain, time-frequency-domain and entropy features, both linear and non-linear. Features extracted from the frequency- or time-frequency domain of the signals are commonly used, derived from either the Fourier or the Wavelet transform [12, 15, 17–19, 31, 33, 39]. Several approaches use a set of different types of features [10, 11, 14, 23, 28, 32, 38] including spectral features such as powers in different frequency bands and wavelet packet coefficients, time domain features such as mean, median and variance of a signal, minimum, maximum and RMS values, and Hjorth complexity parameters. Nonlinear features include for example the Teager energy operator, line length and the Hurst exponent to mention a few and finally, various entropy features have also been used [20, 22]. While many studies rely on a limited number of features extracted from the signals, others search for an optimal combination of features [11, 14, 17, 23–25, 27, 30–32].

Apart from the different features used, existing methods also differ in the type of classification framework [23, 38]. Amongst the blind spectral methods, Neural Networks [12, 14–16, 21, 33, 34, 40] and Decision Trees [17, 18, 32, 35] are frequently employed along with Random Forests [20, 28], Support Vector Machines [11, 24], K-Nearest Neighbors [18, 25], Linear Discriminant Classifiers [22, 27, 30], Hidden Markov Models [13] and Fuzzy Classifiers [19]. For the rule-based approaches the classification has typically been performed in a hierarchical manner by layered models of decision [8, 9, 32]. More recently, a genetic fuzzy inference system has also been used [10].

Although the reported results in the literature cannot all be compared equally due to different evaluation methods, remarkable progress has been made in this field in recent

years, with existing methods reporting an overall scoring agreement to human experts ranging from around 70% to 90%. Figure 8 shows the performance of our algorithm (represented by a pentagram) and other existing methods (each represented by an asterisk) that satisfied three conditions to allow for a more fair comparison: (1) epochs were classified into the five stages suggested by the AASM Manual, (2) the data set consisted of recordings from at least 20 subjects and (3) the test data was independent of the training data. The methods were divided into four groups based on the signals employed for feature extraction. The color of each asterisk on the plot characterizes the type of features used; expert based features (red) represent supervised features based on the rules of the AASM Manual whereas blind spectral features (blue) are unsupervised features that do not require any pre-knowledge about the data. Additionally, some existing studies extracted both expert based and blind spectral features from the PSG signals (magenta).

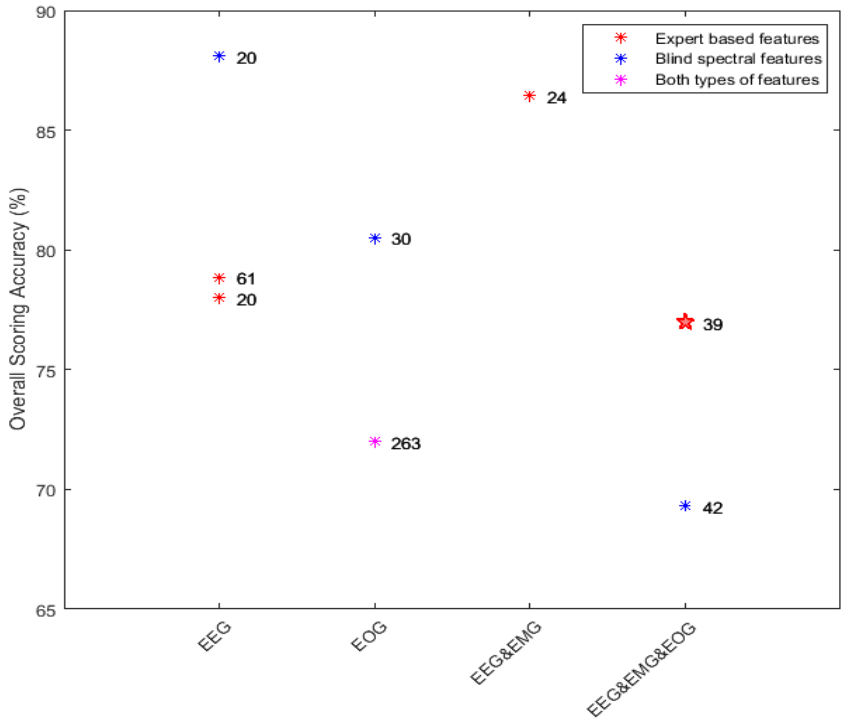

Figure 8. A comparison of the performance of our algorithm (red pentagram) and a few existing scoring methods (asterisks). All studies satisfy certain conditions allowing for a more reasonable comparison.

As Figure 8 shows most studies in the higher range of performance used a smaller data set compared to the studies in the lower range. Using a small data set increases the risk of overfitting, potentially resulting in a high classification accuracy on a particular set of data but generalizing poorly on others. For instance, using only a part of our data set, where we trained the algorithm on 20 subjects as before but tested the performance on only 10 subjects, we were able to increase the overall scoring accuracy of the test set to 81.47%. Reducing the size of the training set as well might increase the accuracy even further. It is evident that many factors can affect the performance of the classifiers and direct comparison between different studies is complicated. The issues of comparing existing methods are further addressed in the Discussion section.

The blind spectral methods are more prevalent in the literature than the rule-based methods and even though the results obtained so far are very promising, limitations still exist. First, for many of the machine learning techniques commonly used, the exact decision procedure remains hidden. Second, most of the numerical classifiers are designed to rely on the same set of features for all sleep stages. Although every feature has an advantage in distinguishing some sleep stages, none of them can discriminate between *all* stages. Third, stage N1 has turned out to be particularly hard to score in the literature due to its similarities to other stages. Stages N1 and REM exhibit similar EEG patterns and moreover since N1 is a transition phase between Wake and the different sleep stages it is easily confused to other stages [37]. In order to overcome this problem many of the scoring algorithms that have been proposed and implemented over the years combine two or more stages into a single stage and thus don't classify into the five stages recommended by the AASM Manual

[15, 16, 19, 27, 31, 35]. Although impressive results have been achieved, these classifiers are not sufficient for replacing human experts.

One drawback of the blind spectral approach is the fact that the AASM scoring criteria are largely based on the identification of "stage specific" wave forms and events such as sleep spindles, K complexes or rapid eye movements rather than the background activity of the signals. As previously mentioned, the blind spectral approaches commonly use features derived from the power spectrum of the signals, but the spectrum is mainly determined by the background signal activities, rather than by these characteristic features that the human scorer is specifically trained to identify as "wave forms that stand out from the background" [4]. Moreover, the AASM Manual has a set of temporal context smoothing rules for epochs where no events are detected. These event-based and smoothing rules, when ignored, can cause low performance of numerical classifiers [36].

Here, we propose an automatic sleep stage scoring method with the goal of improving the speed, reliability, accuracy and cost efficiency of the PSG scoring process compared to the current annotated paradigm. We extracted features from EEG, EMG and EOG signals based on predefined rules of the AASM Manual for Scoring Sleep. The features were computed for 30-second epochs, in either the time or the frequency domain. The most useful features were selected by observing probability distributions for each metric conditioned on the sleep stage, and identifying the features giving the greatest separation between stages. These features were then used as inputs to a likelihood ratio decision tree classifier which assigned each epoch one of five possible stages: N3, N2, N1, REM or Wake.

2 Methods

2.1 Database

A total of 39 healthy adults, 17 males and 22 females, participated in the study (Table 2). The data was randomly split into a "training" and a "test" set with 20 participants in the training group and 19 participants in the test group. The study population consisted of 46.2% Caucasian subjects, 33.3% Asian subjects and 20.5% African American subjects, and participants' age ranged from 18 years to 30 years. All participants were evaluated by a sleep specialist and validated surveys and had no reported sleep disturbances, no apnea by PSG criteria and neither circadian rhythm disorder nor restless leg syndrome. Both "good sleepers" (PSQI score ≤ 5) and "poor sleepers" (PSQI score > 5) as defined by the Pittsburgh Sleep Quality Index (PSQI) [41] were included in the study, although the majority of participants (87.2%) had PSQI score ≤ 5. This study was approved by the JHMI IRB and all participants provided informed consent prior to enrollment.

Table 2. Population statistics.

Subject statistics	Training Set	Test Set
Gender (M / F)	8 / 12	9/ 10
Age (years)	23.8 ± 3.0	24.6 ± 3.4
Ethnicity (Caucasian / Asian / African American)	10 / 5 / 5	8 / 8 / 3
PSQI	2.1 ± 1.7	3.7 ± 2.0
Time in bed (hours)	6.2 ± 0.1	6.2 ± 0.1

2.2 Data Acquisition

The polysomnography was conducted in the Johns Hopkins Clinical Research unit using standardized procedures. The recordings were done for an entire night of sleep and the PSG device model used was the same across all participants. The PSG data, collected for each subject, included six EEG channels collected in a contralateral ear reference montage (F3-A2, F4-A1, C3-A2, C4-A1, O1-A2 and O2-A1) (Figure 9); two EOG channels, one for each eye (right EOG-A2 and left EOG-A2); three EMG channels (chin, right leg and left leg); one ECG channel; respiratory flow and effort, measured as breathing movements by respiratory inductance plethysmography (RIP) belts (thoracic and abdominal); oximeter; thermistor and cannula. The sampling rate of the EEG, EOG, EMG and ECG signals was 500 Hz and 200 Hz for the thermistor and cannula. The respiratory effort was collected at a sampling rate of 10 Hz and the oxygen saturation was sampled at 2 Hz.

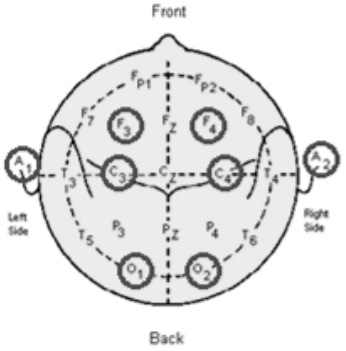

Figure 9. Electrode placements.

2.3 Data Analysis

2.3.1 Human Expert Scoring

All 39 PSG recordings were analyzed by a seasoned licensed and registered sleep technician using the Embla RemLogic sleep diagnostic software. The recordings were visually scored according to the 2007 AASM Manual for Scoring Sleep and Associated Events by assigning one of the five possible sleep stages to every 30-second epoch of the

data. A board-certified sleep specialist reviewed and finalized all recordings, which were conducted and scored by the sleep technician.

2.3.2 Automated Sleep Stage Scoring Algorithm

The proposed automatic sleep scoring system consists of three main steps. The decision rules were inspired by the AASM criteria, and the features were extracted based on the corresponding characteristics of PSG data in the time and frequency domains. These features quantify the sleep scoring rules and translate the human knowledge into metrics that can be used for automatic operation. Furthermore, the model took into account experts in the sleep field with current and past positions on sleep stage guidelines [Allen, R.P., Gamaldo, C.E. & Salas, R.M.E., personal communication]. Secondly, the thresholds for each feature were chosen based on the distributions of feature values at different sleep stages. The classification of sleep stages was performed in a hierarchical manner and thus the probability distributions were drawn using only the remaining epochs after scoring each stage. Thirdly, a likelihood ratio decision tree classifier was utilized to perform the classification and finally a set of temporal contextual smoothing rules was applied on the annotated data, in accordance with what is visually done by experts. The whole scoring process, from feature extraction from raw signals to automatic scoring by the likelihood ratio classifier, was timed for each test subject for comparison to an estimated time of human expert scoring (Appendix B). All data analysis and scoring algorithm implementation was performed using Mathworks MATLAB R2015b.

2.3.2.1 Data Preprocessing

The PSG recordings used for the automatic sleep stage classification included four out of six available EEG channels (F3-A2, F4-A1, C4-A1 and O1-A2), both EOG channels (right and left eye) and all EMG channels (chin, right leg and left leg). In fact, the EEG features were computed from all six channels, but only the channels giving the best separation of sleep stages were used in the analysis. Other PSG signals such as ECG and thermistor were also examined but were not found to provide useful information for our classifier.

The electrophysiological signals were filtered using third order high- and lowpass Butterworth filters. The order of the filters was chosen to best match the appearance of signals as they were displayed in RemLogic. Table 3 summarizes the cutoff frequencies for each signal type, which were selected based on the AASM criteria [4] and filter settings in RemLogic. Artifacts, such as those caused by movements, were not removed from the signals as they were considered important for detecting stage Wake.

Table 3. Filter settings of the PSG signals.

PSG signal	Low-cut frequency (Hz)	High-cut frequency (Hz)
EEG	0.3	35
EOG	0.3	35
EMG	10	-

2.3.2.2 Feature Extraction

The continuous filtered recordings were divided into non-overlapping 30-second epochs for feature extraction. We extracted various features from the signals according to the characteristics of each sleep stage described by the AASM Manual. Normalization of features was employed to reduce the effects of individual differences on classifier performance. For each participant, features that were not normalized inherently by their computation method were normalized using:

$$z_{epoch} = \frac{x_{epoch} - \mu}{\mu} \quad \textbf{(1)}$$

where z_{epoch} is the normalized value for a particular epoch, $x_{epoch}$is the feature value at that epoch and μ is the average feature value across all epochs over the entire night. μ was calculated excluding the highest 5% and lowest 5% of values.

EEG Relative Power in Sleep Frequency Bands

As the different sleep stages have distinct patterns of EEG activity, the relative powers in different frequency bands were used as features in our model. To compute the relative power in a particular frequency band for a given epoch, the power spectrum of the epoch was first obtained using fast Fourier transform (FFT). Then, the total area under the curve (AUC_{tot}) across all frequencies was computed along with the area under the curve corresponding to the frequency band of interest (AUC_{band}). Finally, the relative power was defined as the percentage of area under the curve corresponding to the frequency band of interest:

$$\%\,\mathrm{AUC}_{\mathrm{epoch}} = \frac{AUC_{band}}{AUC_{tot}} \quad \textbf{(2)}$$

The relative power values were normalized for each participant using Eq. (1).

Maximum EMG Energy

Muscular activity is expected to be highest during the Wake stage. Muscular activity is strongly linked to the epoch's energy, where EMG epochs containing high muscular activity have high energy levels and low energy epochs have low muscular activity [42]. Therefore, we expected stage Wake to have the highest energy per epoch. The EMG energy was defined as:

$$Energy = \left[\sum_{n=1}^{N} \frac{\left[X(n) - E[X]\right]^2}{N}\right] \quad \textbf{(3)}$$

where $X(n)$ is the n-th EMG value in an epoch, $E[X]$ is the mean EMG value of the epoch and N is the number of samples per epoch. In order to attempt to better capture the difference in muscular activity per epoch, each 30-second epoch was further divided into six 5 second windows before computing the EMG energy using Eq. (3). The EMG energy was computed in all three EMG channels, hence for each epoch, six EMG energy values were computed in each EMG channel, yielding 18 energy values per epoch. Then, the energy values were normalized for each channel separately using Eq. (1), and finally the maximum energy value per epoch was found and used as a feature for scoring stage Wake.

EOG Correlation

Two types of eye movements can be present in the EOG signal; rapid eye movements (REMs) and slow eye movements (SEMs). SEMs are typically waves around 0.1-0.3 Hz whereas REMs are in the frequency range of 0.3-0.45 Hz [42]. In order to capture the eye movements in the EOG signals we considered two important factors; whether the movements were periodic and whether the movements were conjugate, i.e. the left and right EOG signals were out of phase.

For each epoch, we first computed the cross-correlation and the autocorrelation of the two EOG signals. The cross-correlation of two signals is a measure of the similarity of the signals as a function of the lag of one relative to the other and the autocorrelation of a signal is defined as the cross-correlation of the signal with itself at different points in time. If two signals are out of phase, the peak located at a lag of zero will be negative, with a larger peak corresponding to a greater phase difference between the two signals. In the autocorrelation of a signal, a peak is always present at the zero lag. However, the

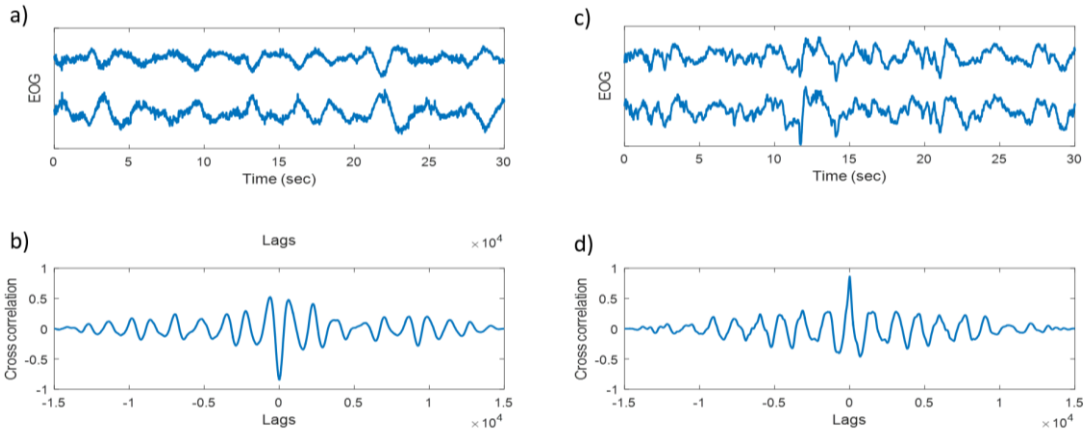

Figure 10. Cross-correlations of two EOG signals. a) EOG signals that are out of phase and b) the corresponding cross-correlation, c) EOG signals that are in phase and d) the corresponding cross-correlation.

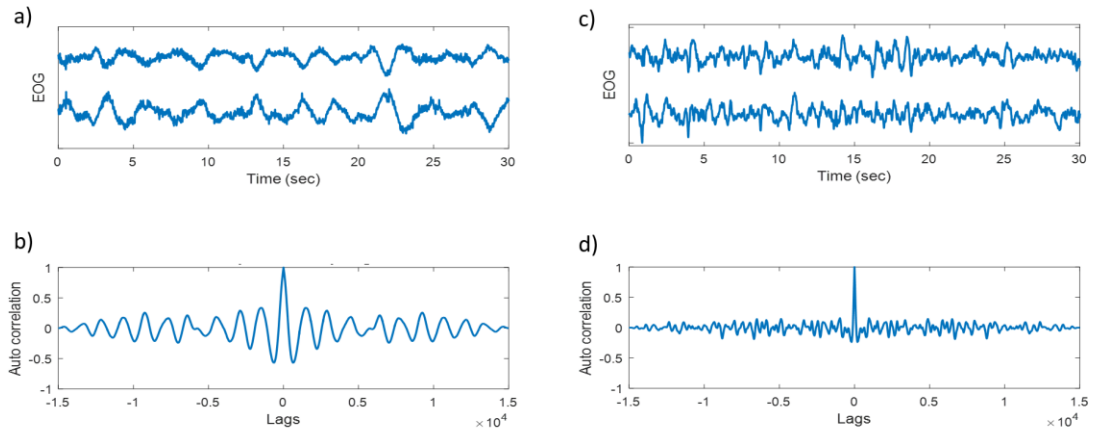

Figure 11. Autocorrelations of an EOG signal. a) A periodic EOG signal and b) the corresponding autocorrelation, c) A non-periodic EOG signal and d) the corresponding autocorrelation.

autocorrelation of a periodic signal is itself periodic and thus, if a signal is periodic, we expect the slope between the peak at zero-lag and the adjacent peak to be smaller compared to the slope for non-periodic signals. Similarly, we expect the average absolute value of the autocorrelation in an epoch to be lower for non-periodic signals compared to periodic signals. Figure 10 shows examples of a) correlated EOG signals that are out of phase and b) the corresponding cross-correlation, and c) correlated EOG signals that are in phase and d) the corresponding cross-correlation. Figure 11 shows examples of a) a periodic EOG signal and b) the corresponding autocorrelation, and c) a non-periodic EOG signal and d) the corresponding autocorrelation.

We computed various features combining these characteristics of the cross- and autocorrelations into a single value for each epoch. All features were computed in four different frequency bands of the EOG signals; 0.3-35 Hz (the frequency range of the EOG signals as suggested by the AASM Manual), 0.1-0.3 Hz (the frequency range of SEMs), 0.3-0.45 Hz (the frequency range of REMs) and 0.1-0.45 Hz (the frequency range

corresponding to both SEMs and REMs). Only the features giving the greatest separation of sleep stages were selected for our analysis. Below is a definition of the correlation features that were computed for each epoch of the EOG signals and used in the classifier:

$$EOG_1^{0.3-35} = \frac{1}{m_{AC}} * sign(CC_{peak}) \quad \textbf{(4)}$$

where m_{AC} is the slope of the adjacent autocorrelation peaks and $CC_{peak}$is the value of the cross-correlation peak located at the zero-lag. $EOG_1^{0.3-35}$ was computed using the EOG signals bandpass filtered in the 0.3-35 Hz range.

$$EOG_2^{0.1-0.45} = \frac{1}{mean(abs(AC)} * CC_{peak} \quad \textbf{(5)}$$

where AC is the autocorrelation of the EOG signals in an epoch and $CC_{peak}$is the value of the cross-correlation peak located at the zero-lag. $EOG_2^{0.1-0.45}$ was computed using the EOG signals bandpass filtered in the 0.1-0.45 Hz range.

$$EOG_3^{0.3-0.45} = (1 - m_{AC}) * CC_{peak} \quad \textbf{(6)}$$

where m_{AC} is the slope of the adjacent autocorrelation peaks and $CC_{peak}$is the value of the cross-correlation peak located at the zero-lag. $EOG_3^{0.3-0.45}$ was computed using the EOG signals bandpass filtered in the 0.3-0.45 Hz range.

Sleep Spindles

Sleep spindles represent the "characteristic" feature for scoring stage N2. We searched for spindles in the F3-A2 and C4-A1 EEG channels using the Wendt algorithm [43], an automatic spindle detector based on the definitions stated by the AASM standard. An advantage of the Wendt algorithm is that no preknowledge of sleep stages is required and thus it is not dependent on the human scoring. The output of the algorithm is a vector of the same length as the input vector of the EEG signal, with ones where spindles are detected and zeros everywhere else. Two features were computed from these output vectors; the Maximum Spindle Duration per epoch in the F3-A2 channel and the Number of Spindles per epoch in the C4-A1 channel. For the Maximum Spindle Duration we found the longest chain of consecutive ones in each epoch. For the Number of Spindles per epoch we first restricted the length of a spindle to 0.5-2 seconds and thereafter counted the chains of consecutive ones that met this criteria in each epoch.

As previously stated, we extracted features from the PSG signals for each stage based on the visual scoring criteria employed by experts. Table 4 lists the features used in the model and their corresponding sleep stages as well as the physiological meaning of each feature. The first feature on the list (EOG_1) was used to split the epochs into two groups of possible stages (N3/N2/N1/REM/Wake vs. N1/REM/Wake only) before assigning each epoch a sleep stage using the other features.

Table 4. The features used in the classifier.

Sleep stage	Quantitative feature	Signal	AASM feature
N1/REM/Wake vs. N3/N2/N1/REM/Wake	$EOG_1^{0.3\text{-}35}$	Right EOG and Left EOG	Eye movements present/absent
Wake	EMG Energy	EMG Chin, Left Leg, Right Leg	Increased EMG activity
	Alpha Power	O1-A2 EEG	Alpha rhythm observed
	Theta Power	O1-A2 EEG	Low theta activity
N1	$EOG_2^{0.1\text{-}0.45}$	Right EOG and Left EOG	Eye movements present
N2	Maximum Spindle Duration	F3-A2 EEG	Spindles present
	Number of Spindles	C4-A1 EEG	Spindles present
	Delta Power	F3-A2 EEG	Moderate to high delta activity
	$EOG_3^{0.3\text{-}0.45}$	Right EOG and Left EOG	Little to no rapid eye movements
N3	Delta Power	F4-A1 EEG	High delta activity
	Beta Power	F3-A2 EEG	Low beta activity

2.3.2.3 Threshold Determination

The decision thresholds for each feature were set based on the probability distributions of feature values, conditioned on the sleep stage. The PSG data was divided into a training and a test set, with whole night recordings from 20 subjects in the training group and 19 subjects in the test group, and the thresholds were chosen using only the training set data. For each feature we used the expert annotations to draw five probability distributions, one for each stage (see section 3.1), and then chose a threshold attempting to optimize the number of epochs detected for the conditioned sleep stage with minimum decision error. This was often located close to intersection of distributions where the probability of the conditioned sleep stage became higher than for any other sleep stage. Two features and their sleep stage probability distributions, along with the corresponding classification thresholds are shown for clarification in Figure 12. All feature probability distributions can be seen in section 3.1

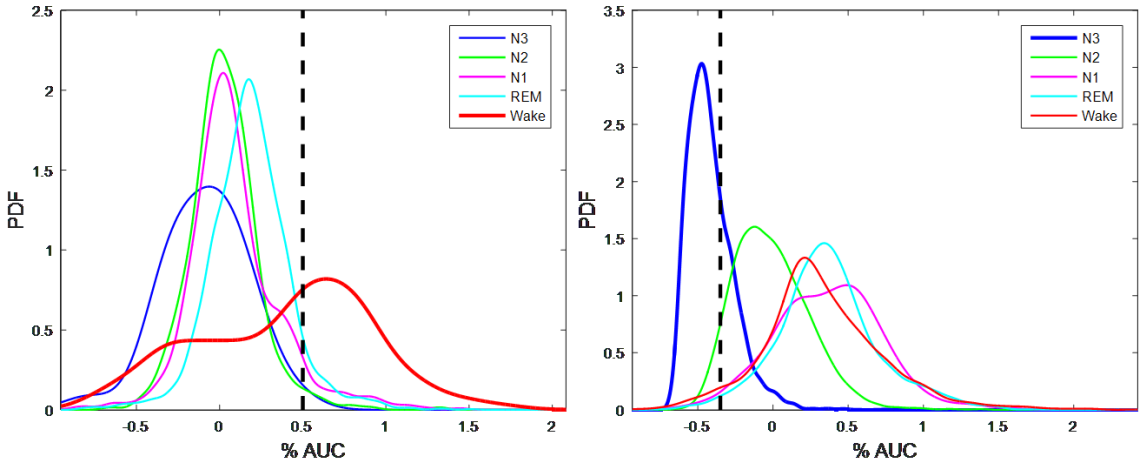

Figure 12. Probability density functions and the corresponding decision thresholds for relative Alpha power, a feature for stage Wake (left), and relative Beta power, a feature for stage N3 (right).

2.3.2.4 Automatic Sleep Stage Classification

The classification process of the likelihood ratio decision tree classifier (Figure 13) comprised five steps. First, epochs were divided into two groups based on the auto- and cross-correlations of the EOG signals (Eq. (4)). If the $EOG_1^{0.3\text{-}35}$ value was below a certain threshold, indicating that little to no eye movement was present, the epoch was assigned to group 1, and if it was above the threshold the epoch was assigned to group 2. In group 1 all five sleep stages were possible, whereas epochs in group 2 could only be scored as N1, REM or Wake.

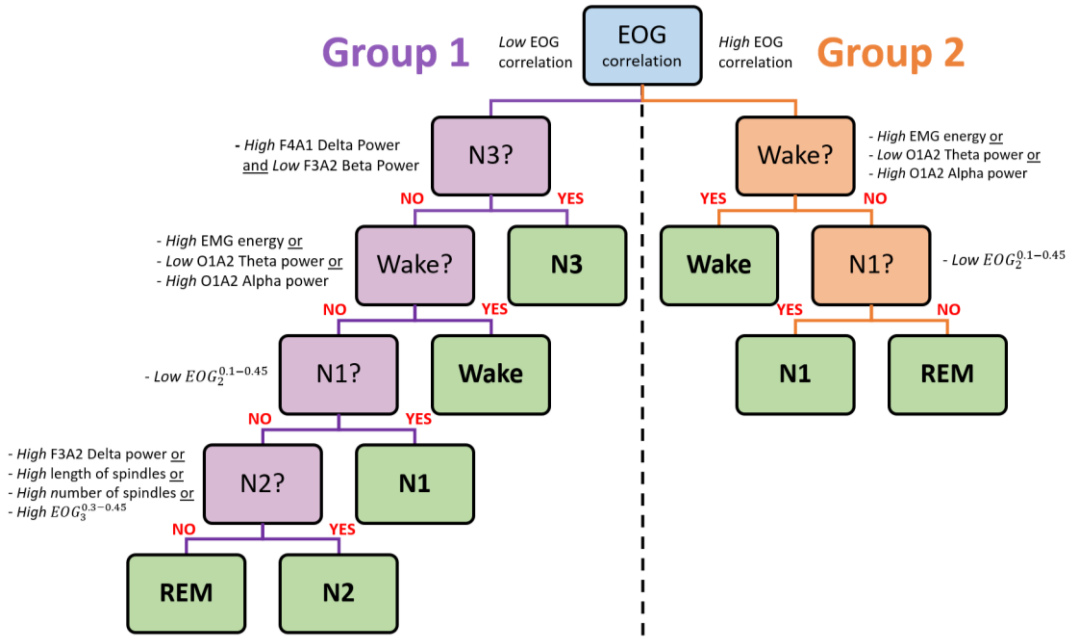

Figure 13. A flowchart of the automatic scoring process of the likelihood ratio decision tree classifier.

In the second step, all epochs belonging to group 1 were assigned a sleep stage. The sleep stages were scored in the following order: N3, Wake, N1, N2 and REM. The scoring order was selected based on the discriminating ability of the features, with stages that were easier to detect scored first. Once an epoch was assigned a sleep stage it was excluded from the scoring of the remaining stages, that is, only unscored epochs were considered when scoring each stage. In our analysis, no quantitative feature was found to be informative enough to discriminate stage REM from the other stages. Consequently, after scoring Wake and the three non-REM stages, all remaining unscored epochs were assigned to stage REM.

The AASM Manual has a number of rules that recommend considering the neighboring epochs for the scoring of a current epoch under certain circumstances. Thus, when performing sleep stage scoring, an expert may refer to the neighboring epochs in addition to the current epoch to make decisions. Additionally, sleep is a continuous process and alternating between different sleep stages every 30 seconds is highly unlikely. Therefore, a smoothing process considering the temporal contextual information was applied after scoring the epochs in group 1. These contextual smoothing rules refer to the relationship between epochs prior to and after the current epoch. Specifically, let A, B and C represent the possible stages (N3, N2, N1, REM or Wake). Then, three consecutive epochs of A, B, A were replaced with A, A, A and four consecutive epochs of A, B, B, A or A, B, C, A were replaced with A, A, A, A.

The fourth step consisted of scoring the epochs in group 2. As stated previously, epochs in group 2 were only scored as N1, REM or Wake since eye movement activity was high in those epochs. In this group, stage Wake was scored first, followed by N1. As for group 1, the remaining unscored epochs were set as stage REM.

Lastly, the same set of contextual smoothing rules were applied to the epochs in group 2. Additionally, in compliance with the AASM Scoring Manual, the first epoch scored as any other stage than Wake was set as N1.

3 Results

The performance of the proposed algorithm was evaluated by comparing the agreement between the automatic classification and the human expert scoring, which served as the gold standard. The training set was first used to determine the optimal thresholds for each feature as well as the most effective scoring order of sleep stages that provided the maximum classification accuracy. Then, the performance of the likelihood ratio classifier was evaluated using the test set. We report the performance of the classifier for both training and test data. Interestingly, the overall accuracy of the test data set is very similar to that of the train data set, indicating the robustness of the proposed algorithm on different data sets.

3.1 Quantitative Feature Distributions

For each feature, the expert annotations were used to draw five probability distributions, one for each stage. The classification of sleep stages was performed in a hierarchical manner and thus the probability distributions were drawn using only the remaining epochs after scoring each stage. The feature probability distributions below are presented in the same order as they appear in the scoring algorithm.

Feature for dividing epochs into Groups 1 and 2

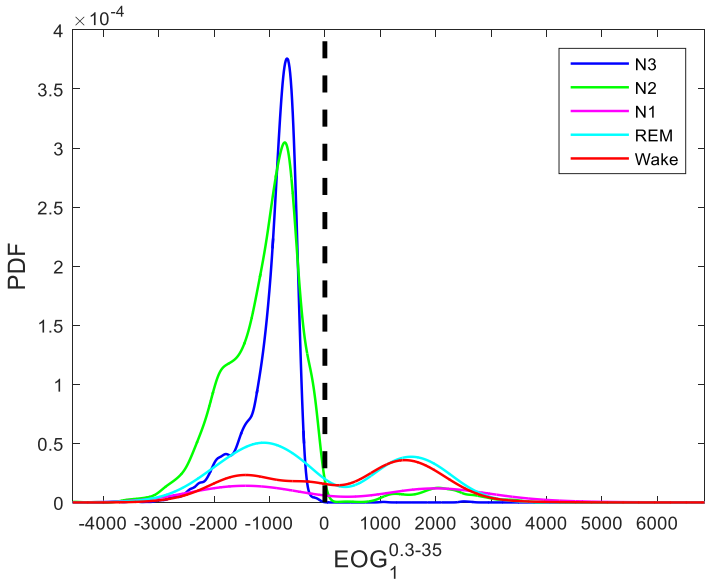

Figure 14. Probability distributions of each sleep stage for $EOG_1^{0.3\text{-}35}$. The values on the x-axis represent normalized feature values. Epochs with $EOG_1^{0.3\text{-}35} \leq 0$ were assigned to group 1 and epochs with $EOG_1^{0.3\text{-}35} > 0$ were assigned to group 2.

Features for Scoring Epochs in Group 1

N3 Features

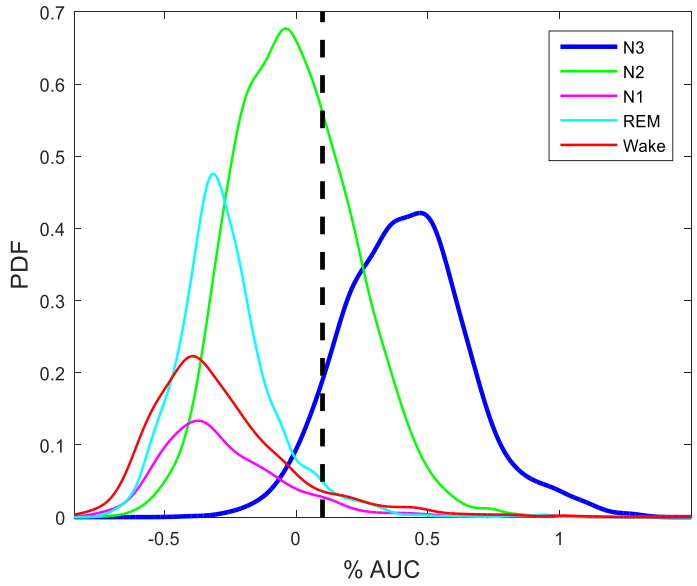

Figure 15. Probability distributions of each sleep stage for Delta Power in group 1. The values on the x-axis represent normalized feature values. The probability curve for N3 is shown in blue. Epochs with Delta Power ≥ 0.1 and Beta Power ≤ -0.35 were scored as N3.

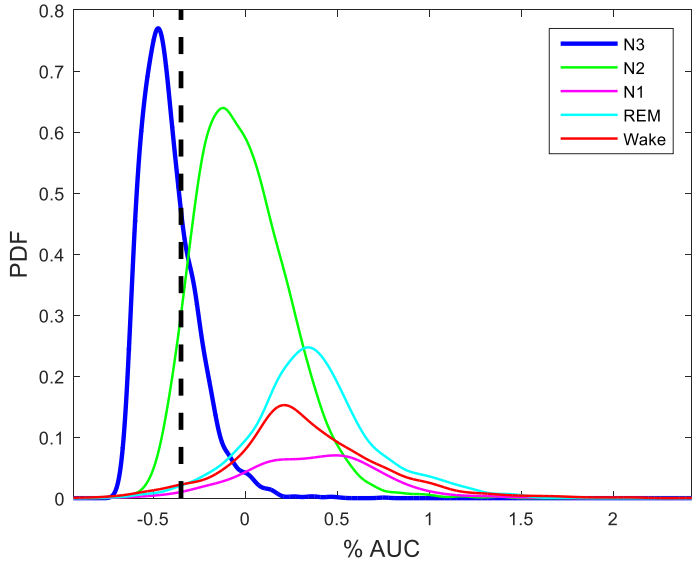

Figure 16. Probability distributions of each sleep stage for Beta Power in group 1. The values on the x-axis represent normalized feature values. The probability curve for N3 is shown in blue. Epochs with Delta Power ≥ 0.1 and Beta Power ≤ -0.35 were scored as N3.

Wake Features

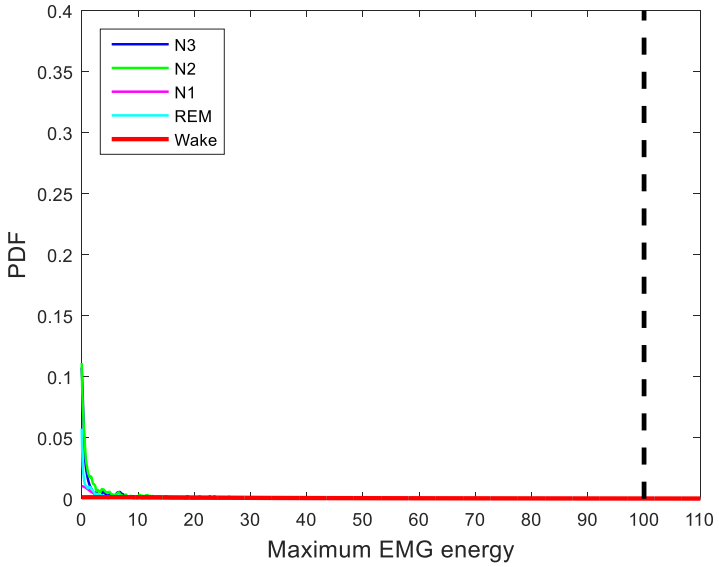

Figure 17. Probability distributions of each sleep stage for Maximum EMG energy in group 1. The values on the x-axis represent normalized feature values. The EMG Energy values ranged from 0 to 4500 but here we have zoomed in on the x-axis for clarification purposes. The probability curve for Wake is shown in red and the probability curves of stages that have already been scored are shown faded. Epochs with EMG Energy ≥ 100 were scored as Wake.

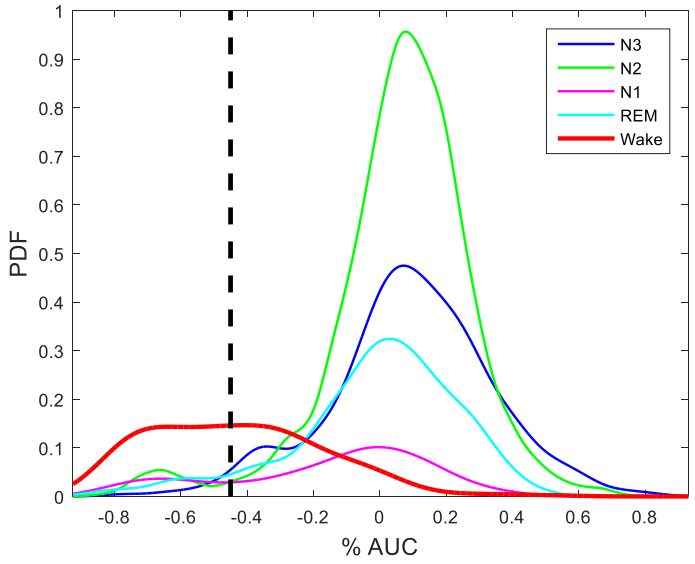

Figure 18. Probability distributions of each sleep stage for Theta Power in group 1. The values on the x-axis represent normalized feature values. The probability curve for Wake is shown in red and the probability curves of stages that have already been scored are shown faded. Epochs with Theta Power ≤ -0.45 were scored as Wake.

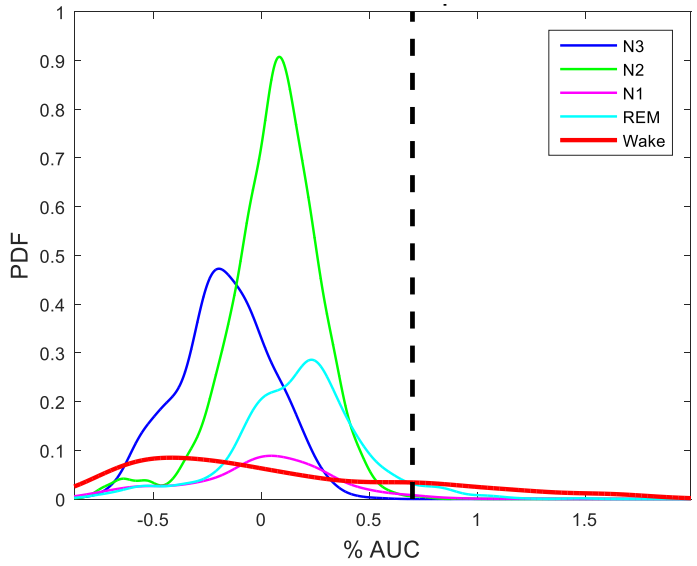

Figure 19. Probability distributions of each sleep stage for Alpha Power in group 1. The values on the x-axis represent normalized feature values. The probability curve for Wake is shown in red and the probability curves of stages that have already been scored are shown faded. Epochs with Alpha Power ≥ 0.7 were scored as Wake.

N1 Feature

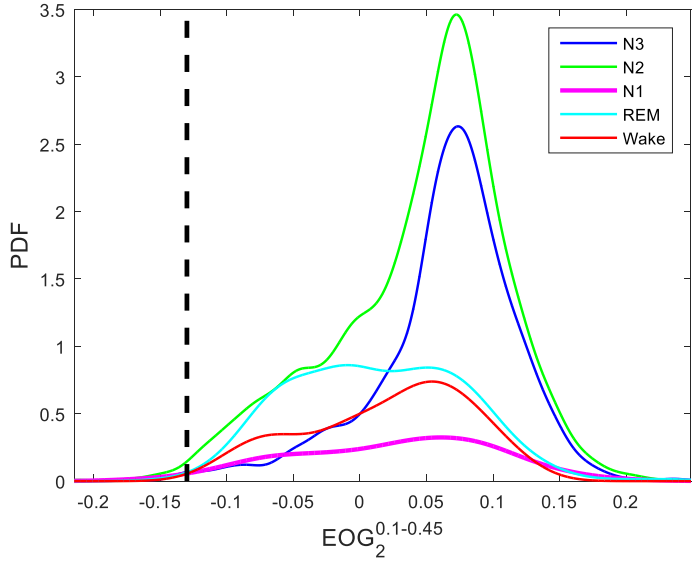

Figure 20. Probability distributions of each sleep stage for $EOG_2^{0.1\text{-}0.45}$ in group 1. The values on the x-axis represent normalized feature values. The probability curve for N1 is shown in magenta and the probability curves of stages that have already been scored are shown faded. Epochs with $EOG_2^{0.1\text{-}0.45} \leq -0.13$ were scored as N1.

N2 Features

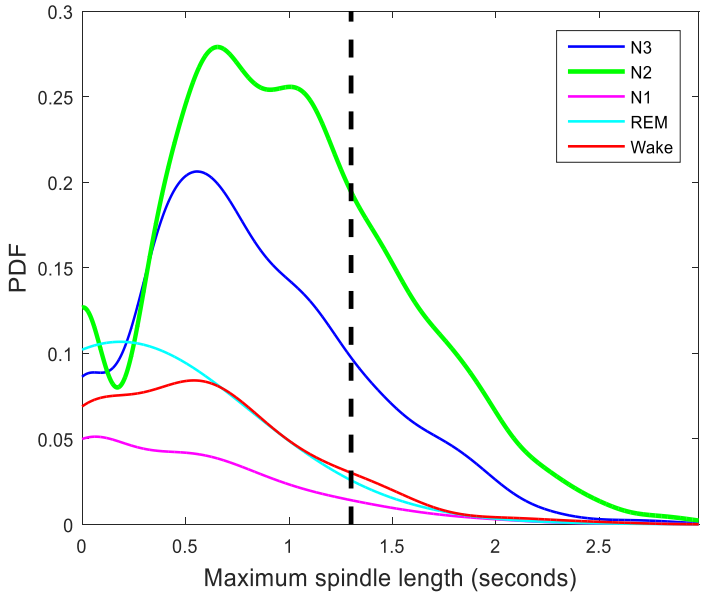

Figure 21. Probability distributions of each sleep stage for Maximum Spindle Duration in group 1. The probability curve for N2 is shown in green and the probability curves of stages that have already been scored are shown faded. Epochs with Maximum Spindle Duration ≥ 1.3 were scored as N2.

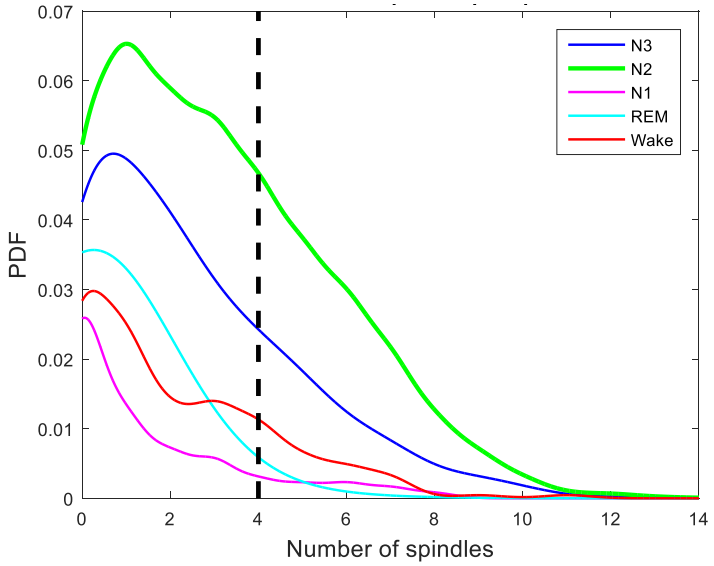

Figure 22. Probability distributions of each sleep stage for Number of Spindles in group 1. The probability curve for N2 is shown in green and the probability curves of stages that have already been scored are shown faded. Epochs with Number of Spindles ≥ 4 were scored as N2.

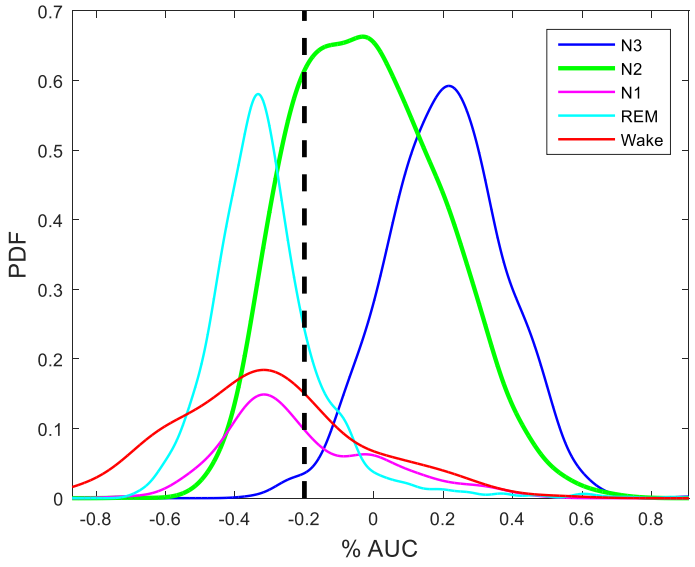

Figure 23. Probability distributions of each sleep stage for Delta Power in group 1. The values on the x-axis represent normalized feature values. The probability curve for N2 is shown in green and the probability curves of stages that have already been scored are shown faded. Epochs with Delta Power ≥ -0.2 were scored as N2.

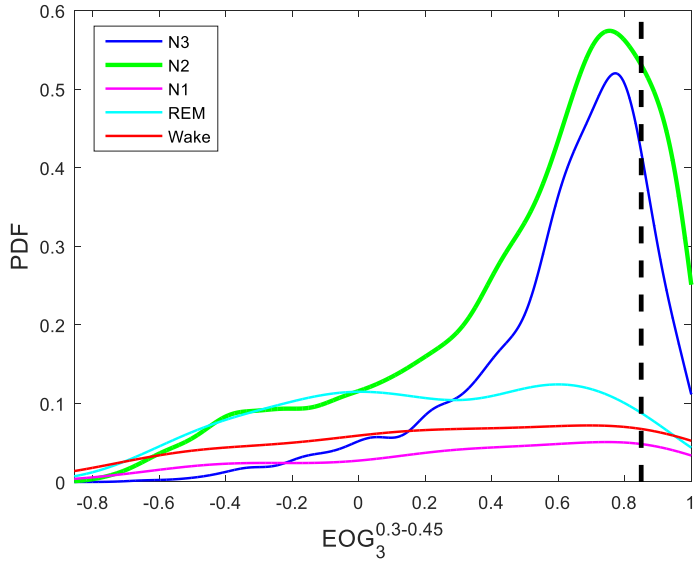

Figure 24. Probability distributions of each sleep stage for $EOG_3^{0.3\text{-}0.45}$ in group 1. The values on the x-axis represent normalized feature values. The probability curve for N2 is shown in green and the probability curves of stages that have already been scored are shown faded. Epochs with $EOG_3^{0.3\text{-}0.45} \geq 0.85$ were scored as N2.

Features for Scoring Epochs in Group 2

Wake Features

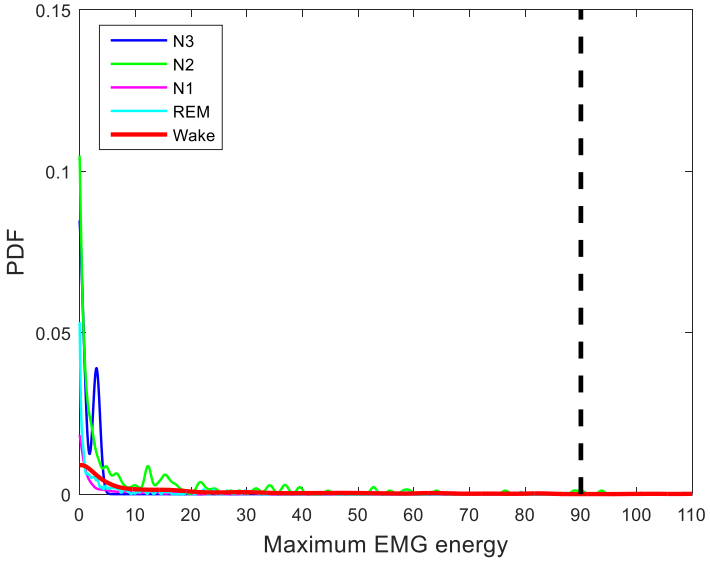

Figure 25. Probability distributions of each sleep stage for Maximum EMG energy in group 2. The values on the x-axis represent normalized feature values. The EMG Energy values ranged from 0 to 6000 but here we have zoomed in on the x-axis for clarification purposes. The probability curve for Wake is shown in red and the probability curves of stages that have already been scored are shown faded. Epochs with EMG Energy ≥ 90 were scored as Wake.

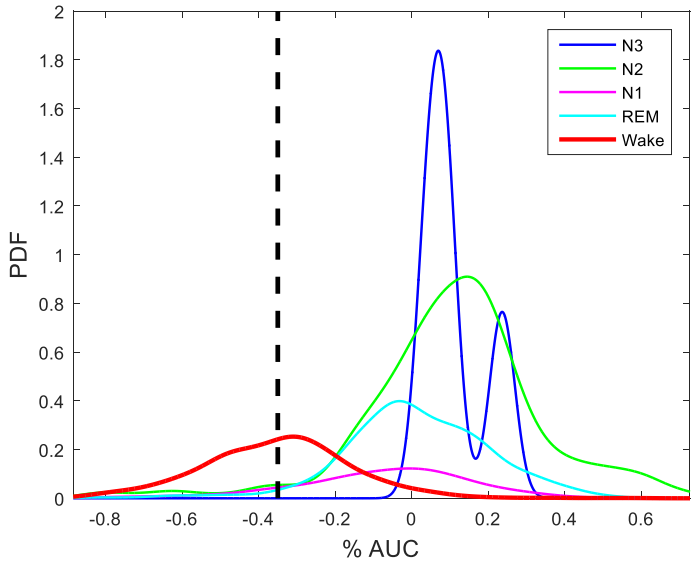

Figure 26. Probability distributions of each sleep stage for Theta Power in group 2. The values on the x-axis represent normalized feature values. The probability curve for Wake is shown in red and the probability curves of stages that have already been scored are shown faded. Epochs with Theta Power ≤ -0.35 were scored as Wake.

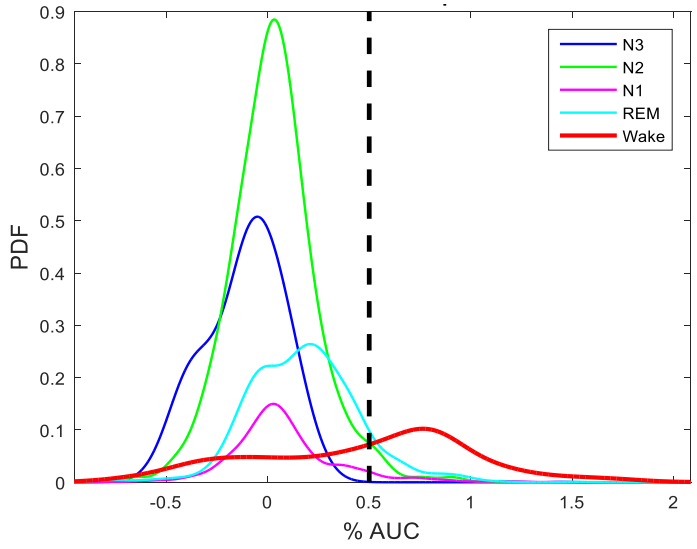

Figure 27. Probability distributions of each sleep stage for Alpha Power in group 2. The values on the x-axis represent normalized feature values. The probability curve for Wake is shown in red and the probability curves of stages that have already been scored are shown faded. Epochs with Alpha Power ≥ 0.5 were scored as Wake.

N1 Feature

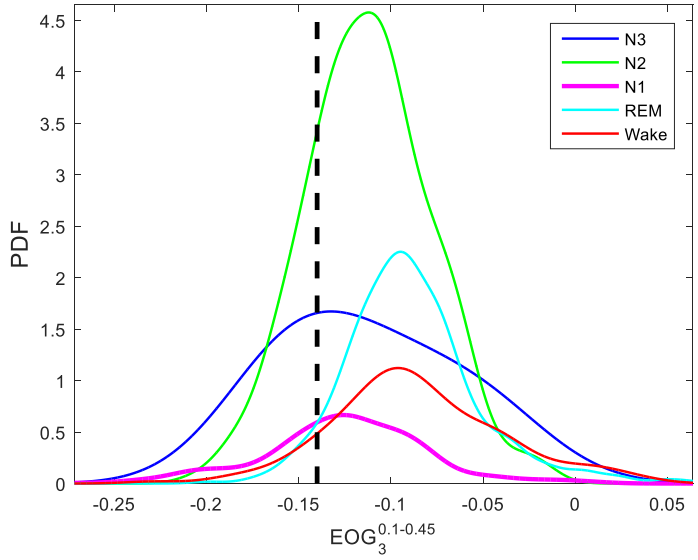

Figure 28. Probability distributions of each sleep stage for $EOG_2^{0.1\text{-}0.45}$ in group 2. The values on the x-axis represent normalized feature values. The probability curve for N1 is shown in magenta and the probability curves of stages that have already been scored are shown faded. Epochs with $EOG_2^{0.1\text{-}0.45} \leq -0.14$ were scored as N1.

3.2 Training Set Results

Of the 15,257 epochs in the training set, the best performance of the algorithm resulted in 11,700 correctly classified epochs, or an overall scoring accuracy of 76.69%. The best scoring performance was obtained for sleep stage N3, with 90.29% scoring accuracy and a total of 75.44% of N3 epochs captured. In contrast, stage N1 turned out to be the most difficult stage to detect, with 15.32% of N1 epochs captured and a scoring accuracy of 57.03% (Table 5). Figure 29 shows the confusion matrix after using the classifier on the training data set. The rows of the matrix represent the actual scoring by the human expert and the columns of the matrix represent the predicted scoring by the proposed algorithm. The values are the percentage of epochs belonging to the stage scored by the expert (indicated by the rows) that were classified by our algorithm as the stage indicated by the columns. The diagonal elements are shown in bold and represent the percentage of epochs where the automatic classifier was in agreement with the human expert for each sleep stage. For N1, most misclassifications occurred between the N1-N2 pair, followed by N1-REM and N1-Wake. Other commonly misclassified pairs were N3-N2, REM-N2 and Wake-N2. The remaining pairs all had misclassification rates below 10%.

Table 5. Scoring results of the training set. The scoring accuracy and epochs captured are reported for each sleep stage along with the overall scoring accuracy of the training data set.

Sleep Stage	Scoring Accuracy (%)	Epochs Captured (%)
Wake	70.71	79.26
N1	57.03	15.32
N2	74.46	86.53
N3	90.29	75.44
REM	72.52	76.91
Overall Scoring Accuracy	**76.69%**	

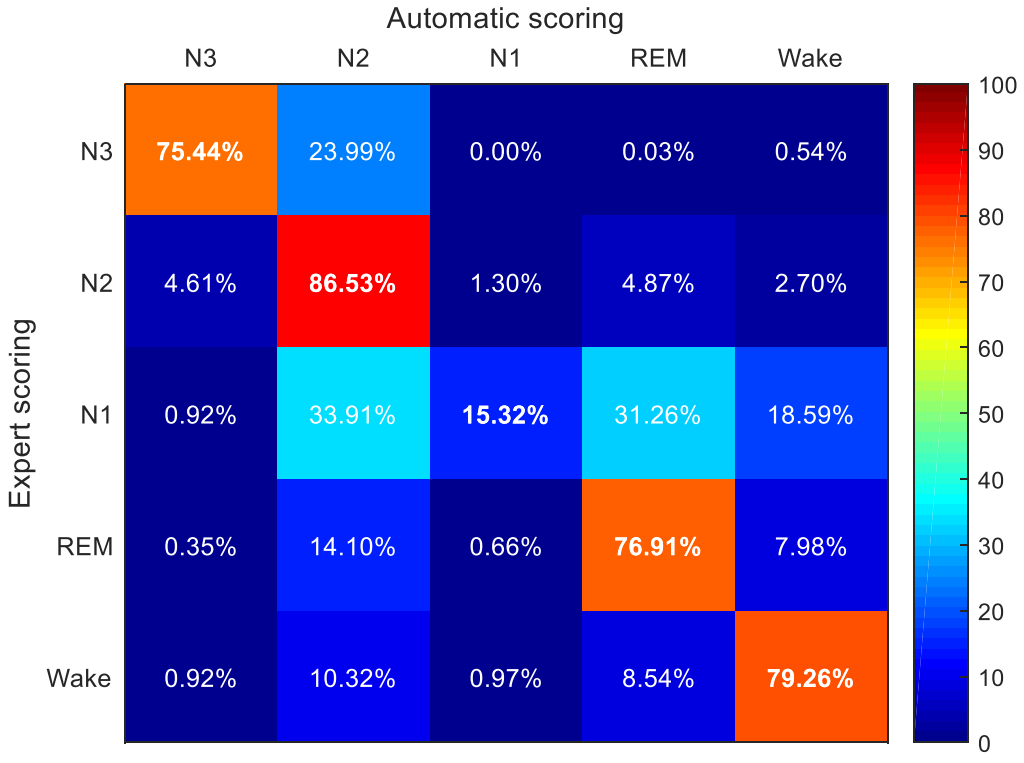

Figure 29. Confusion matrix for the automatic scoring algorithm using the training data set.

3.3 Test Set Results

The overall scoring accuracy of the test set was slightly higher than for the training set, with 11,031 epochs out of 14,332 correctly classified, resulting in an overall scoring accuracy of 76.97%. The highest scoring accuracy of a single subject was 89.37% and the lowest accuracy was 61.59%. Hypnograms for all test subjects along with the associated scoring accuracy can be found in Appendix A. The accuracy pattern was analogous to the training set, with the highest scoring accuracy obtained for stage N3 and the lowest for stage N1. The highest percentage of epochs correctly classified was obtained for stage N2, followed closely by Wake. As Table 6 shows the scoring accuracy as well as the number of epochs captured were improved for stage Wake using the test set compared to the training set, with a scoring accuracy of 78.31% and correctly scoring 86.97% of all epochs scored as Wake by the expert. The scoring accuracy of stage N3 was higher than for the training set, or 91.02%, but with a lower number of captured epochs. Conversely, the number of epochs in agreement with the human scorer was higher for both N2 and REM but with a slightly lower scoring accuracy compared to the training set. N1 continued to be the hardest stage to score with a scoring accuracy below 50% and far less epochs captured compared to the other stages.

Figure 30 shows the confusion matrix after scoring the test set using the proposed scoring algorithm. As for the training set, N1 was the stage most often confused with other stages. For N1, most misclassifications occurred between the N1-REM pair, followed by N1-N2 and N1-Wake. Other commonly misclassified pairs were N3-N2 and REM-N2. The remaining pairs all had misclassification rates below 10%.

Table 6. Scoring results of the test set. The scoring accuracy and epochs captured are reported for each sleep stage along with the overall scoring accuracy of the test data set.

Sleep Stage	Scoring Accuracy (%)	Epochs Captured (%)
Wake	78.31	86.97
N1	40.93	10.85
N2	74.24	89.10
N3	90.55	65.18
REM	72.38	77.41
Overall Scoring Accuracy	**76.97%**	

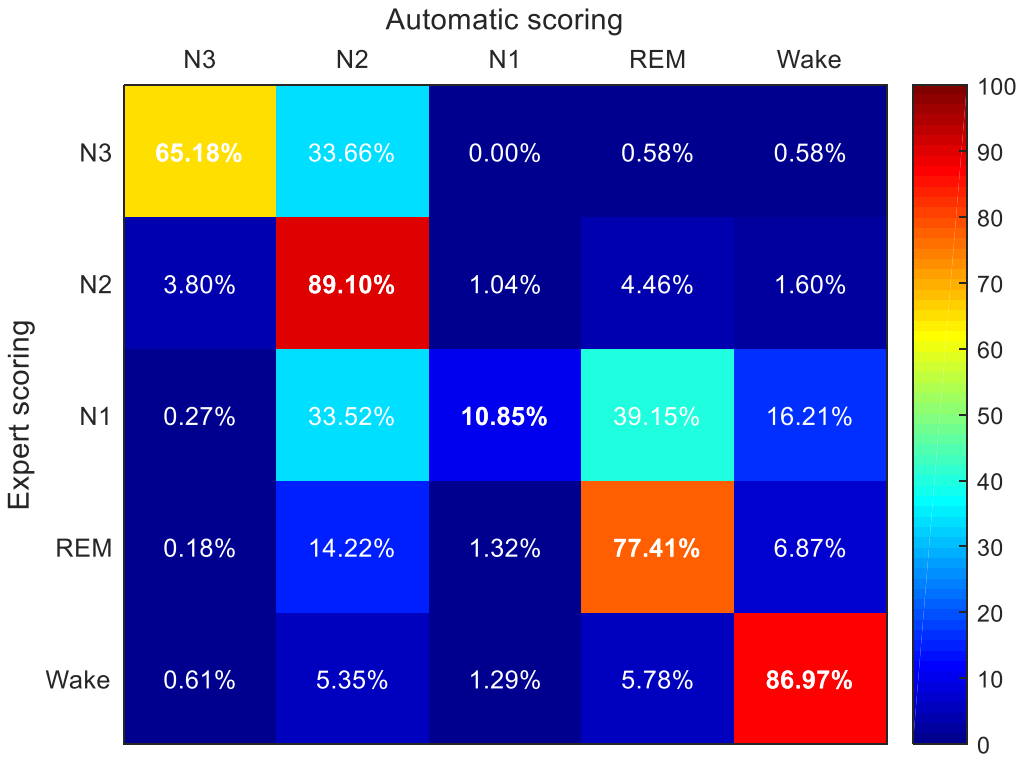

Figure 30. Confusion matrix for the automatic scoring algorithm using the test data set.

Wake versus Sleep Analysis

For some diagnosis and analysis of PSG data, a full sleep stage scoring of all five stages may not be necessary. By combining the four sleep stages (N3, N2, N1 and REM) into a single stage of Sleep, we report an overall scoring accuracy of 95.79% on the test set (Table 7). Figure 31 shows a 2-by-2 confusion matrix of scoring Sleep versus Wake using the test set. Out of all manually scored Sleep epochs, 96.91% were detected with a scoring accuracy of 98.31% by the likelihood ratio classifier. The scoring accuracy of stage Wake remains unchanged from the previous analysis of all five stages, or 78.31%, with a total of 86.97% of all Wake epochs correctly scored.

Table 7. Sleep vs. Wake scoring results of the test set. The scoring accuracy and epochs captured are reported for stage Sleep and stage Wake along with the overall scoring accuracy of Sleep vs. Wake.

Sleep Stage	Scoring Accuracy (%)	Epochs Captured (%)
Sleep	98.31	96.91
Wake	78.31	86.97
Overall Scoring Accuracy	**95.79%**	

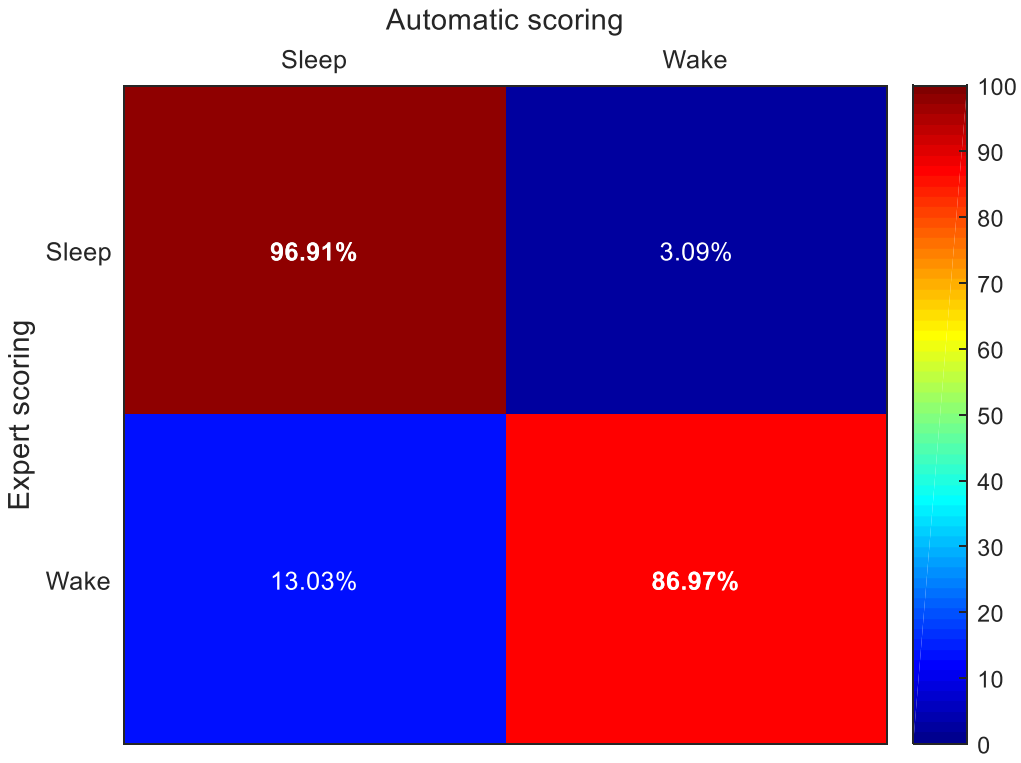

Figure 31. Sleep vs. Wake confusion matrix for the automatic scoring algorithm using the test data set.

4 Discussion

At present, the standard procedure of PSG data analysis is heavily dependent upon human factors and involves a costly and laborious process of sleep stage scoring by sleep specialists which can result in poor inter-scorer reliability. Here, a new system for automatic sleep stage scoring of PSG data has been proposed. The algorithm was trained on a set of 20 subjects and its performance was evaluated on a test set of 19 subjects, with an overall scoring accuracy of 76.97% on the test set.

An important limiting factor of the visual scoring by human experts is the amount of time it takes to score each study and consequently the high expense of the procedure. Not only does it contribute to high operating costs of sleep centers but is also expensive in terms of valuable expert time. Furthermore, in the fast paced environment of the modern world, pressure and time restraints can impair the quality of the sleep stage scoring. A seasoned registered sleep technologist at the Johns Hopkins Sleep Center takes around 30 minutes to 1.5 hours on average to stage a full night sleep study across all physiologic channels monitored. In comparison, the run-time of our algorithm was 32.5 ± 1.9 seconds on average for feature extraction and scoring of a full night sleep recording of a single subject (See Appendix B for the run-times of all test subjects). It is thus clear that automating the scoring process can greatly increase the efficiency of sleep stage scoring by reducing the time and cost of the procedure.

The performances of existing automatic sleep scoring methods show a high degree of variability, with the agreement between human scoring and automated classifiers ranging from around 70% to 93% [9, 16]. While there are reports of higher classification accuracies

in the literature than reported here and many of the results reported are promising, not all methods can be compared equally. First of all, there is a great variability in the size and quality of the data sets used, with many studies applying their algorithms on small sets of data (less than 20 subjects). Here, a data set of 39 subjects was used in order to ensure robust results, but using a sufficient amount of recordings to train and test the classifier is important to ensure the reliability of the algorithm and consistent performance, not only when applied on different individuals but also across various data sets. In addition, some studies do not partition their data into train and test sets, and thus they may likely not obtain a reliable evaluation of their algorithm performance. Of those studies that do split into train and test sets, different methods of training and testing have also been employed. Some studies have for example trained the algorithm on a part of the data from each subject and tested on the remaining data for each subject, while other studies have split the subjects into two groups and trained the algorithm using the data from one group and then evaluated the performance using the data in the second group.

The types of the recorded data sets differ across existing methods as well. Stage N2 constitutes around 45-55% of total sleep, putting a large weight on N2 and possibly yielding biased results. While many studies apply their algorithms on whole night recordings, others have sampled the same number of epochs from each sleep stage in order to avoid this imbalance between stages. Additionally, many of the studies obtaining classification accuracies in the high end of the reported range combine the most commonly confused stages into a single stage and therefore do not classify the recordings into the five sleep stages as suggested by the AASM Manual.

Finally, inter-rater variability in visual sleep scoring is also a limitation that can affect the results reported using automatic methods. Some studies have compared the algorithm performance to the sleep stage scoring of a single human expert only, while others are comparing to records scored by more than one scorer. Moreover, a number of studies only report the performance on epochs where more than one scorers agreed on the epoch's sleep stage during visual scoring, yielding a "cleaner" set of data for training and testing the classifier.

The annotated inter-scorer reliability amongst sleep experts and seasoned technicians has been reported to be only about 82% [5]. Here, the decision thresholds of each feature were based on the human scored data, which in fact might not be as discriminative because of the poor inter-scorer reliability. Inter-individual variability in the recordings can also affect the classifier performance and the parameters retained are likely not as distinctive as the ones chosen visually by human experts.

Thus, based on these aforementioned factors and limitations noted in the current annotated and automated scoring methods, it is clear that direct comparison between different studies is complicated and the performance of the proposed algorithm represents several desirable and superlative features. Furthermore, the scoring accuracy of the test set was almost identical to (and even slightly higher than) the scoring accuracy of the training set, suggesting the robustness of performance on different data sets. However, it is worth noting that (similar to most of the published reliability studies for previously developed automated staging algorithms) the performance of the classifier was only tested on PSG data from healthy individuals without any sleep disturbances documented. The scoring of data from individuals suffering from sleep disorders poses a greater challenge to both human experts

and computerized scoring procedures. Hence, the proposed algorithm may be expected to provide less accurate results under such circumstances.

Şen et al. [23] sought to identify the most effective features and classification algorithm for automatic sleep stage scoring. By examining various feature extraction methods and comparing five frequently used classifiers they achieved an overall classification accuracy of 98.02%. Even though the results are impressive and this might appear to be an outstanding classification performance, it is likely not the most suitable method. The reason is that the accuracy of the classifier cannot be expected to exceed the "gold standard" human inter-rater reliability without the possibility of overfitting, resulting in a worse performance on a new data set or when compared to other human scorers.

In our analysis, stage N2 was the most correctly classified stage, followed by Wake and REM. 89.10% of the epochs scored as N2 by the expert were correctly scored by the algorithm and out of all epochs scored as N2 by the classifier, 74.24% were in agreement with the human scorer. The results for Wake were similar, 86.97% of Wake epochs scored by the expert were detected by the classifier with an accuracy of 78.31%. The least amount of misclassifications occurred for N3, with 90.55% of the epochs scored as N3 by the algorithm correctly classified. However, the detection rate of N3 was a little lower, or 65.18%. Stage N1 was by far the sleep stage most commonly confused with other stages, with only 79 epochs out of 728 correctly scored. The majority of the misclassification errors is likely due to absence of characteristic sleep stage features or the presence of multiple or overlapping features within a single epoch. Furthermore, it is possible that, in some cases, our quantitative features may not sufficiently capture differences in the signal patterns. Signal interference is another problem, where the EMG or EOG signals pick up

the activity of the EEG signals or vice versa, making it difficult to capture the sleep stage characteristics. Moreover, some epochs include periods of a transition from one stage to another, where it is challenging, even for a human expert, to make a decision.

These results are similar to performances of existing sleep stage classification systems, with stage N1 recurrently being the most misclassified sleep stage [9, 13, 14, 17, 18, 20–22, 30, 32, 33]. Moreover, the results are in accordance with reported agreement rates amongst human experts [44]. Generally, the best agreement is achieved with stages N2, Wake and REM. Disagreement with the scoring of stage N1 includes scoring of N2, Wake and REM, and N3 is most frequently confused with stage N2.

The most challenging stage for feature extraction was the REM stage. Studies have shown that the transition between wakefulness and sleep covers some electrophysiological elements which are common to REM sleep, with an increase in REM-like EEG activity after alpha attenuation, right before definite occurrence of sleep spindles [37]. Moreover, according to the AASM Manual, K complexes or sleep spindles are sometimes present in REM epochs, especially in the first REM period of the night. This is also evident when observing the sleep stage probability distributions for the number of spindles per epoch (Figure 22). The distribution for stage N2 overlapped completely with the distribution for REM (and in fact the distributions for all other stages) and as the figure shows, the presence of spindles was fairly common in epochs scored as REM. Epochs with no rapid eye movements present can still be scored as REM, as long as the chin muscle tone remains low and the EEG is in the low mixed frequency range, which in isolation can make distinction between N1 and REM sleep quite difficult. Finally the AASM Manual has no rules to deal specifically with transitions between N1 and REM. As a result, stage REM

shares spectral similarities with Wake, N1 and N2 and thus can get easily confused with other stages.

In our analysis, 39.15% of expert scored N1 epochs were scored as REM by the automatic classifier, and roughly 16% were scored as Wake. The definition of sleep stages by the AASM Manual and the sleep literature [32] show that N1 and REM exhibit similar EEG patterns and since N1 is a transition phase between Wake and the other sleep stages, Wake and N1 share certain spectral similarities as well, making it difficult to distinguish between those stages. Identifying the transition from Wake to N1 is challenging, particularly in subjects with an attenuated alpha activity during Wake or if a subject demonstrates a large amount of tonic REM sleep (REM sleep periods without evidence of rapid eye movements), since stage N1 is scored when EEG theta activity predominates over alpha activity. Even among human experts, the agreement rate for N1 is far below that observed for the other stages, which might be explained by the fact that 10-20% of the population generate little or no alpha rhythm [4], complicating the determination of sleep onset. Furthermore, N1 constitutes only 2-5% of total sleep [2], and thus the data available for training the classifier was very limited.

As expected, most disagreements occur with the scoring of adjacent sleep stages and epochs where a transition from one sleep stage to another takes place. Other pairs of sleep stages frequently misclassified were N2-N3 and N1-N2, but almost 34% of the expert scored N1 epochs were scored as N2. The discrimination of N2 from N1 depends to a great extent on the detection of K Complexes in the EEG signals, associated or not with arousals, body movements in the EMG signals and slow eye movements in the EOG signals, all of which can be difficult to capture, for example because of interference between the different

channels. The misclassification between N3 and N2 can be, at least partly, related to the potential persistence of sleep spindles in stage N3.

Additionally, some epochs do not clearly represent any sleep stage. Moreover, transitions between two adjacent sleep stages may occur in the middle of an epoch or may last longer than 30 seconds during which it is difficult, even for a human expert, to be certain of his/her decision. These epochs represent a problem both in the context of computerized and human expert scoring. An interval containing transitions is only partially classifiable as it is a mixture of stages and thus there is no sleep stage that can serve as a rule for training the classifier. According to the AASM Manual the epoch stage should be assigned based on the predominant stage of the epoch. This means that if more than 50% of the epoch contains characteristics for a certain stage, the epoch should be scored as that stage. This has led us to the consideration of whether 30 seconds are necessarily the ideal length of an epoch or if shorter epochs, for example 10 seconds, can possibly improve the scoring accuracy.

Finally, for some diagnosis and analysis of PSG data, a full sleep stage scoring of all five stages may not be necessary. The temporal distribution of sleep versus wakefulness over the night as well as the total sleep time provide for example valuable information for clinicians and having the option of obtaining this information automatically can save a lot of time and effort otherwise spent manually looking at and scoring all 30-second epochs over the entire night. By combining the four sleep stages (N3, N2, N1 and REM) into a single stage of Sleep, thereby overcoming the issues of the most misclassified transition pairs, we were able to accurately predict the sleep-wake architecture and report an overall scoring accuracy of 95.79% on the test set.

5 Conclusions and Future Work

5.1 Conclusions

In this study we developed an automatic sleep stage scoring method that closely follows the AASM Manual for Scoring Sleep, with the goal of improving the speed, reliability, accuracy and cost efficiency of the PSG scoring process. We extracted features from the physiological recordings of the PSG, based on predefined rules according to the AASM guidelines. The features were computed for 30-second consecutive epochs, in either the time or the frequency domain and the ones giving the greatest separation between sleep stages were identified. These features were then used as inputs to a likelihood ratio decision tree classifier which assigned each epoch one of five possible stages; N3, N2, N1, REM or Wake.

The algorithm was trained and tested on PSG data from 39 individuals with no sleep disturbances. The overall scoring accuracy was 76.97% on the test set. Some of the stages, such as stage N3 have more distinctive characteristics and thus yield a higher per-stage scoring accuracy, whereas other stages, particularly N1, get more easily confused, resulting in lower per-stage accuracies. As expected, most of the disagreements with the human expert occurred with the scoring of adjacent sleep stages. Although this accuracy may at first seem low, the variability in inter-scorer staging, particularly across different sleep stages, has also been reported with human inter-scorer studies [6, 44]. Thus, it is likely that the stages that the automatic scoring tool classified inaccurately may be the same sleep stages that currently contribute to inter-scorer variability within the present annotated paradigm.

The results suggest that the automatic classification is highly consistent with human sleep scoring and that the error in the algorithm is likely due to the ambiguous boundaries between adjacent sleep stages inherent within the current scoring guidelines. There remains an ongoing challenge for human experts to make decisions in certain situations, and thus the disagreements may be more reflective of the "gold standard" approach rather than an insufficiency of the computational procedures.

We conclude that an automatic classification algorithm based on a likelihood ratio classifier, and importantly, using features extracted from the AASM Manual, can to a large extent reproduce the judgment of a sleep scoring human expert. Therefore, we see this tool as assisting sleep scorers to speed up their process and providing a way to diagnose sleeping disorders in a more robust, quantitative and ultimately cost-effective manner.

5.2 Future Work

Future work will aim at further improving the performance of the proposed sleep stage scoring method. Removal of artifacts is an important first step that will yield a cleaner set of data and may therefore increase the overall scoring accuracy. Additionally, further exploration of the PSG signals and identifying useful features based on other rules of the AASM Manual that have not yet been employed is essential to improve the detection of individual sleep stages, particularly N1, N3 and REM. Furthermore, it is desirable to extend these results in young healthy individuals to older healthy individuals and to the more challenging cases of patients with sleep problems.

In the current model, no feature was found to be informative enough to discriminate between REM and the other stages and thus stage REM was assigned to all remaining unscored epochs after scoring the other four stages. In our algorithm, the presence of eye movements in other stages than REM (and the absence of rapid eye movements during tonic REM periods) hindered the utilization of EOG signals for scoring REM. However, the rapid eye blinks that can occur during Wake appear in the EOG recordings as synchronized, positive polarity wave forms across both eyes, whereas the rapid eye movements that occur during REM sleep are characterized by divergent – opposite polarity wave forms across the eyes. This may help to differentiate the rapid eye movements of REM sleep from the eye blinks of wakefulness and thus a further analysis of eye movements in the EOG signals is desirable.

When performing sleep stage scoring, an expert constantly takes into account the contextual information such as the sleep stages of neighboring epochs or the duration of certain phenomena. Thus, we expect improvements of our current contextual smoothing rules, for example by taking into account sleep stage transition patterns described in the AASM Manual and restricting the types of allowed transitions, as well as capturing the duration of particular features to further enhance the agreement with human experts.

Finally, a potential augmentation of the current algorithm is the addition of confidence estimation to the decision of each sleep stage. This would provide the option of manually scoring the epochs with lowest confidence, nevertheless saving experts a great amount of time and effort otherwise spent scoring the whole night recordings. However, it is our hope that our automatic sleep stage scoring algorithm will eventually contribute to improving the efficiency of sleep stage scoring by providing a fast, reliable and accurate alternative

to the current gold standard manual scoring, ultimately replacing the tedious work performed by human experts.

6 References

6.1 Appendices

Appendix A: Test Set Hypnograms

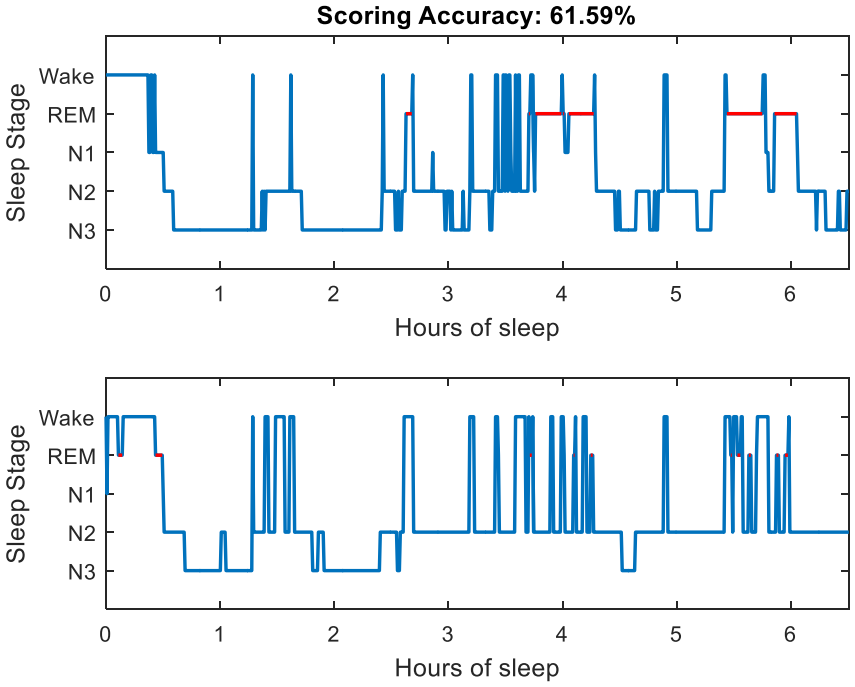

Figure A-1. Hypnograms and scoring accuracy for test subject 1. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

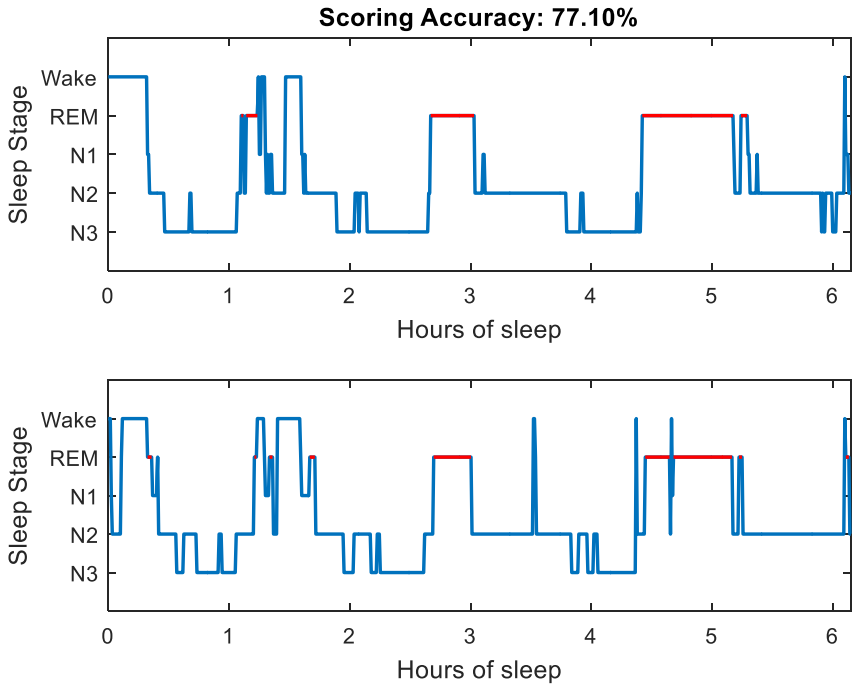

Figure A-2. Hypnograms and scoring accuracy for test subject 2. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

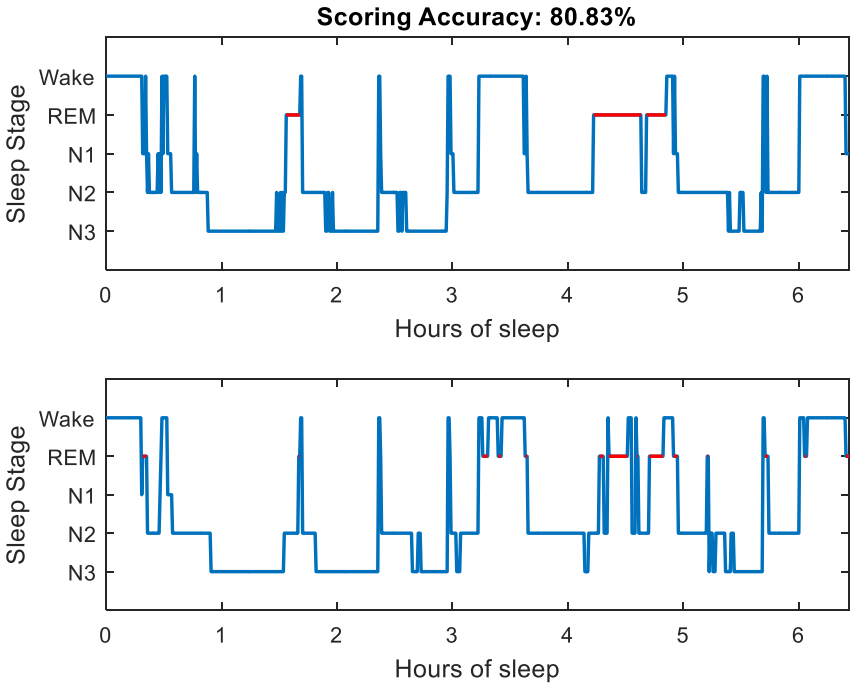

Figure A-3. Hypnograms and scoring accuracy for test subject 3. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

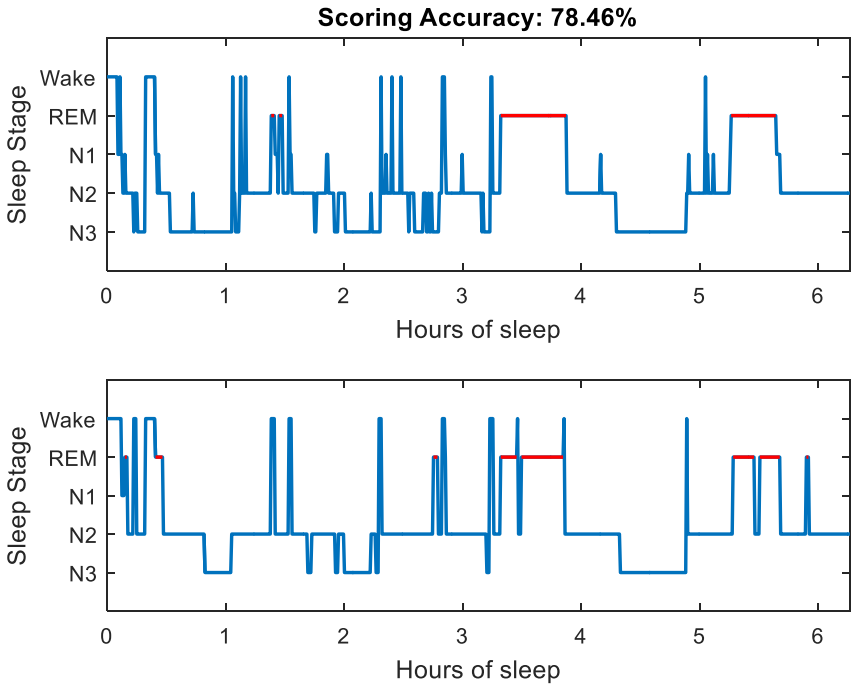

Figure A-4. Hypnograms and scoring accuracy for test subject 4. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

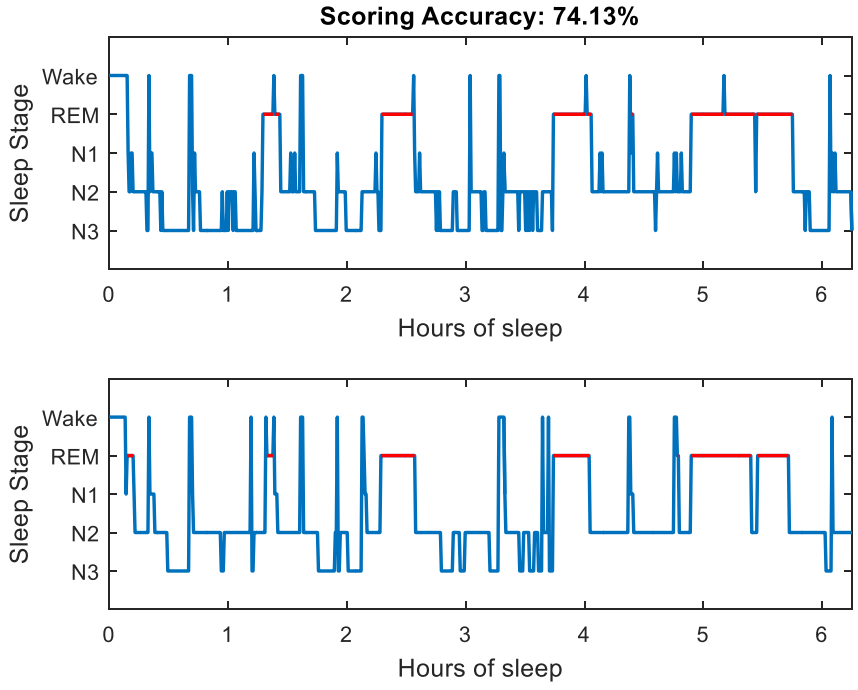

Figure A-5. Hypnograms and scoring accuracy for test subject 5. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

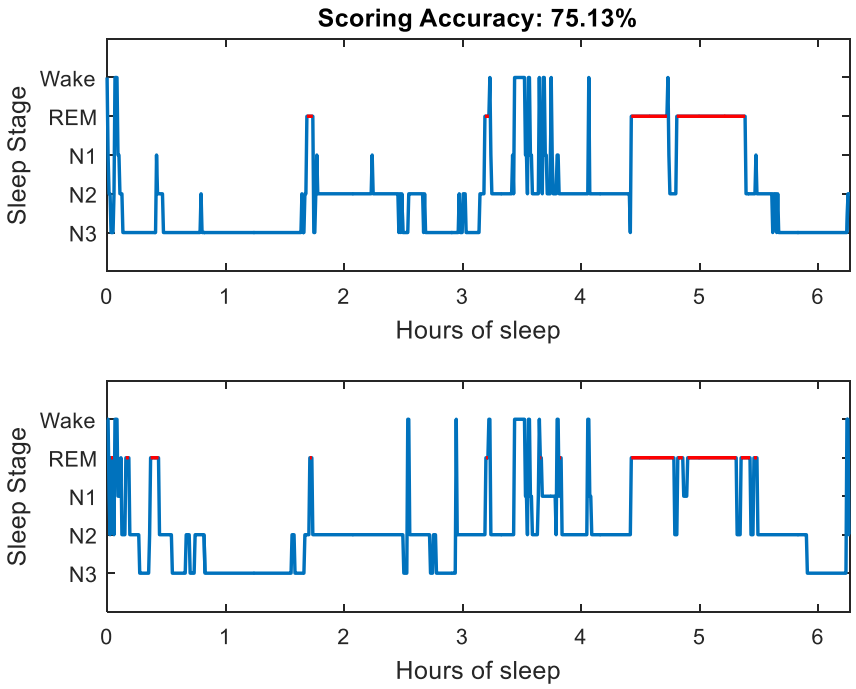

Figure A-6. Hypnograms and scoring accuracy for test subject 6. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

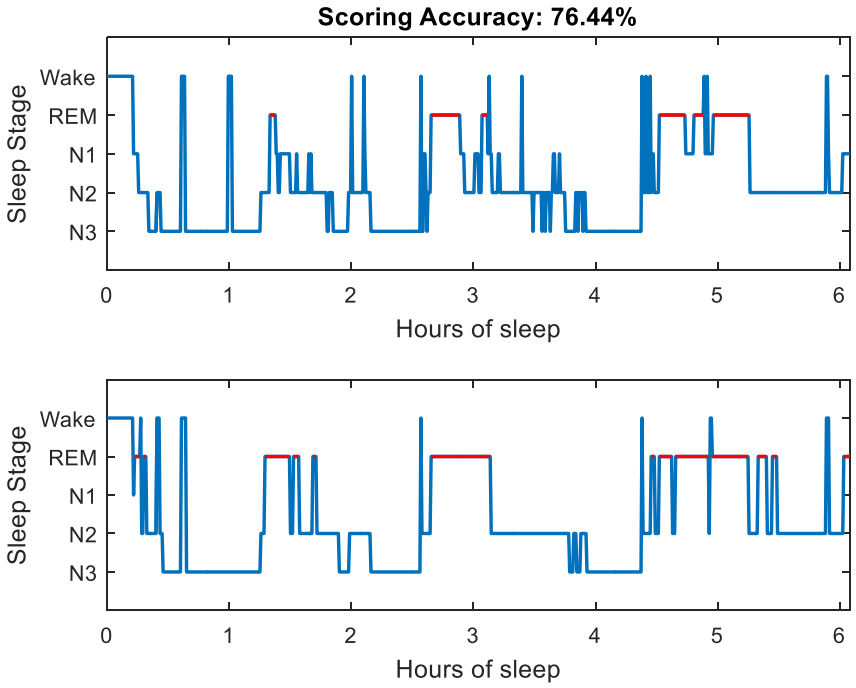

Figure A-7. Hypnograms and scoring accuracy for test subject 7. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

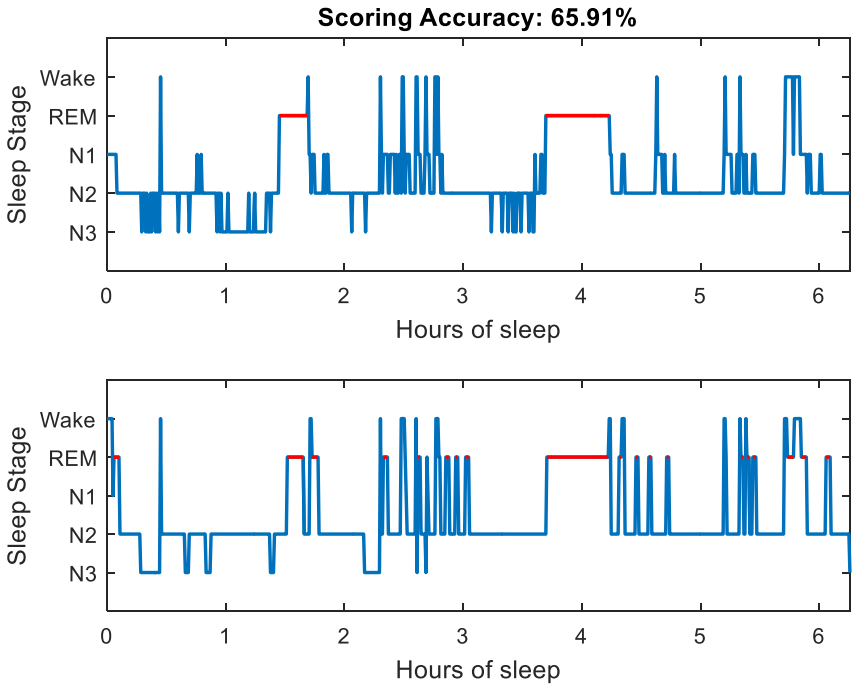

Figure A-8. Hypnograms and scoring accuracy for test subject 8. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

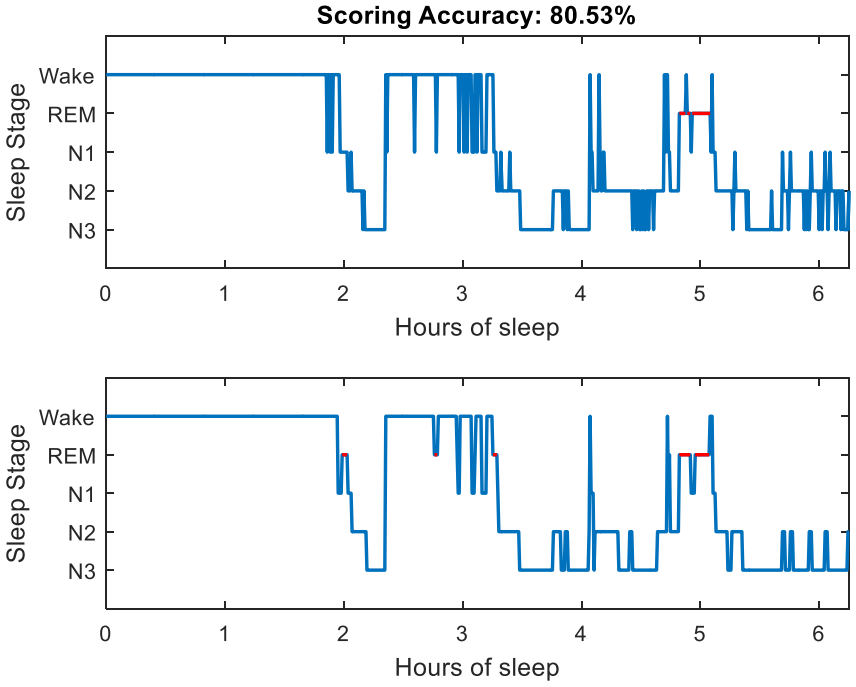

Figure A-9. Hypnograms and scoring accuracy for test subject 9. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

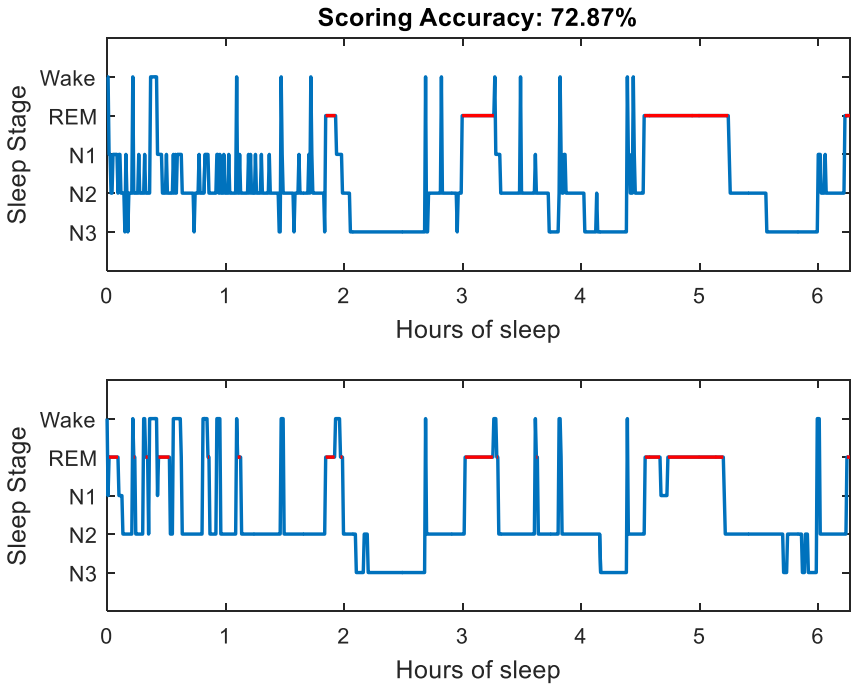

Figure A-10. Hypnograms and scoring accuracy for test subject 10. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

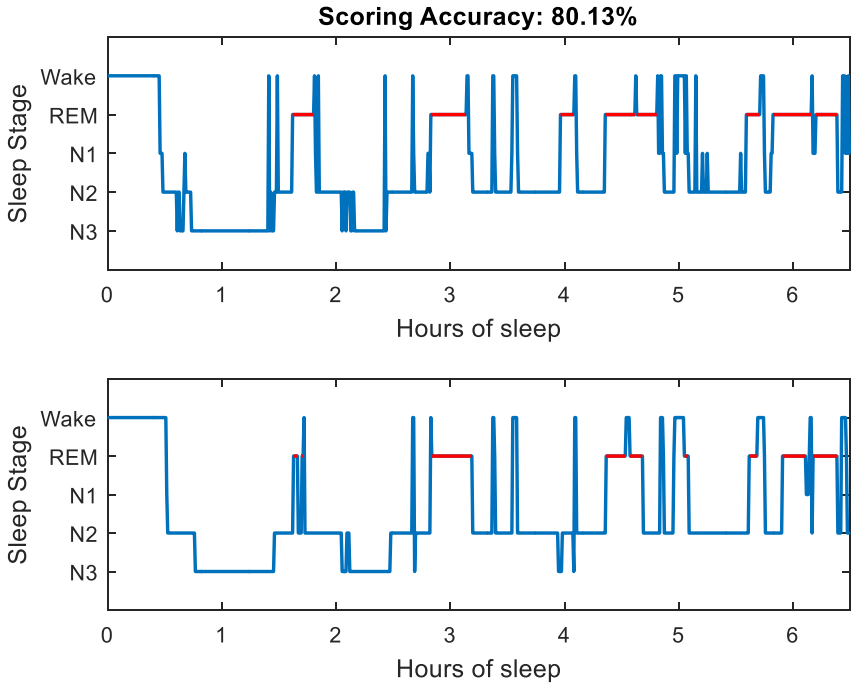

Figure A-11. Hypnograms and scoring accuracy for test subject 11. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

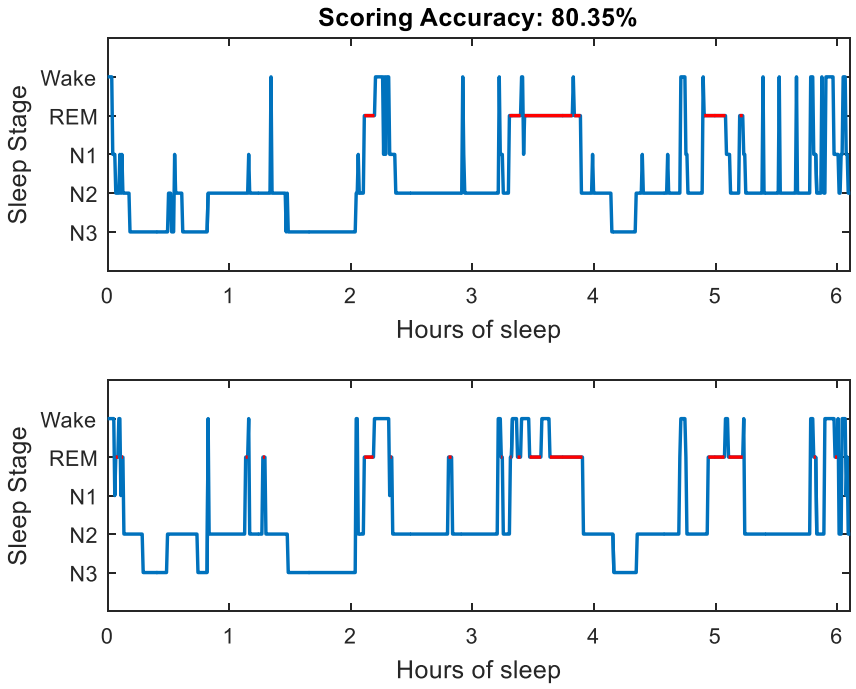

Figure A-12. Hypnograms and scoring accuracy for test subject 12. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

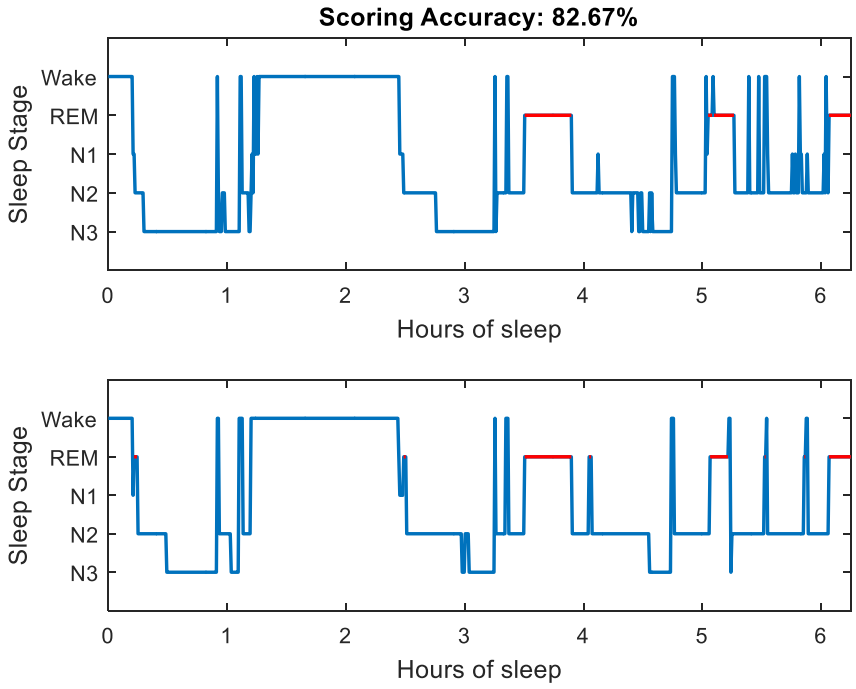

Figure A-13. Hypnograms and scoring accuracy for test subject 13. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

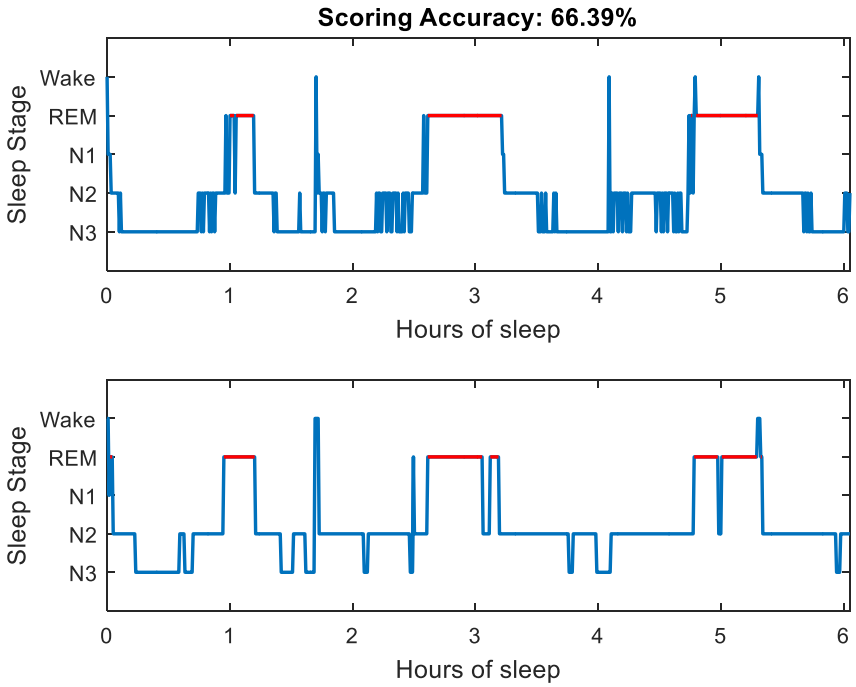

Figure A-14. Hypnograms and scoring accuracy for test subject 14. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

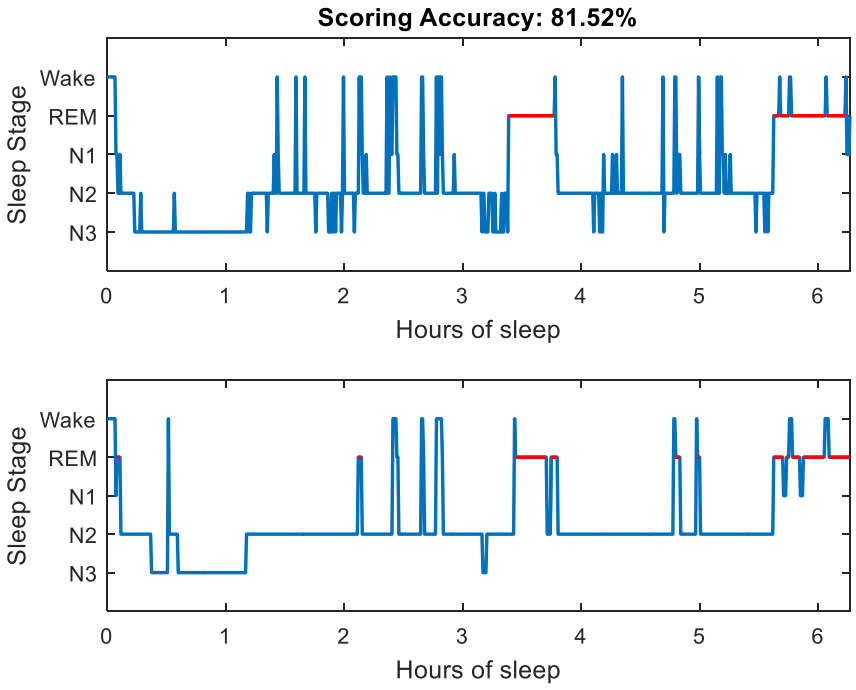

Figure A-15. Hypnograms and scoring accuracy for test subject 15. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

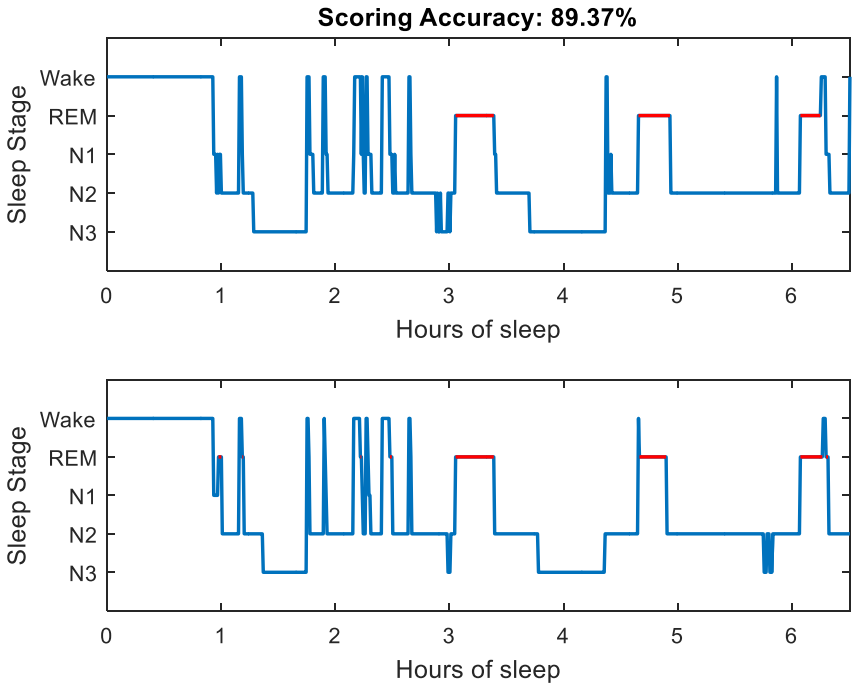

Figure A-16. Hypnograms and scoring accuracy for test subject 16. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

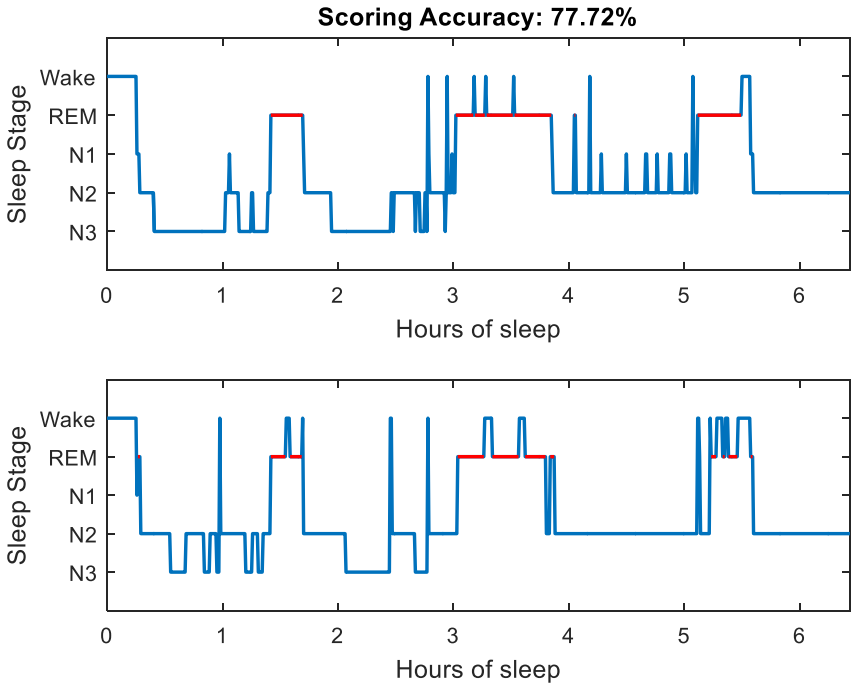

Figure A-17. Hypnograms and scoring accuracy for test subject 17. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

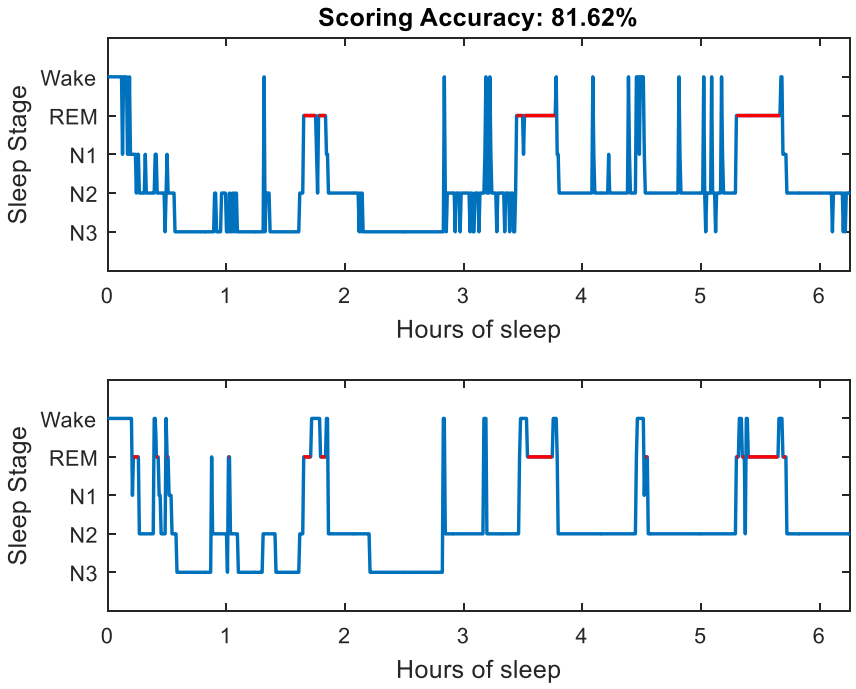

Figure A-18. Hypnograms and scoring accuracy for test subject 18. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

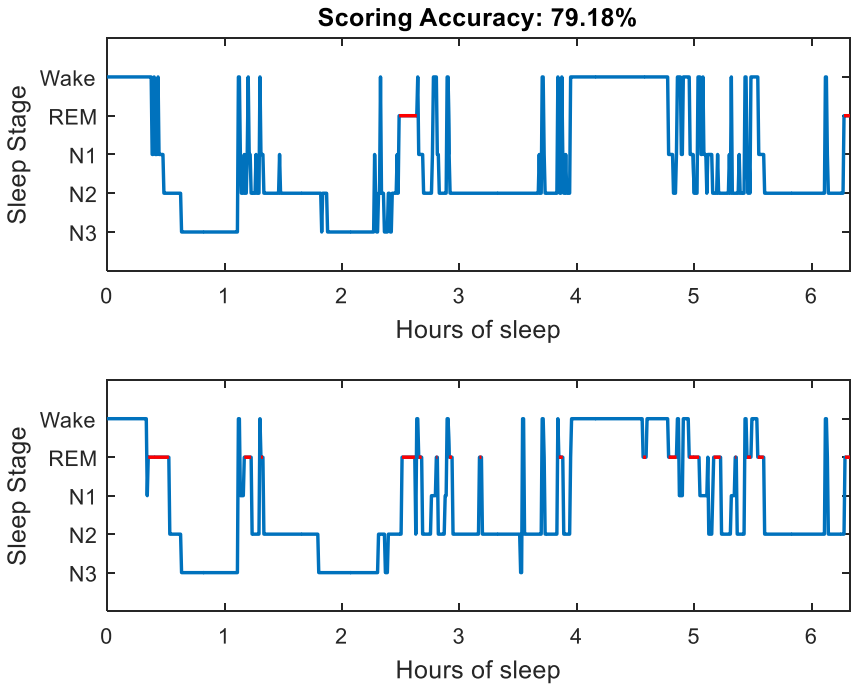

Figure A-19. Hypnograms and scoring accuracy for test subject 19. A comparison of the hypnogram scored by the human expert (top) and the hypnogram generated by the algorithm (bottom).

Appendix B: Test Set Run-Times

Table A-1. The scoring run-times of all test subjects. Scoring run-time represents the time it took to extract features from the PSG signals and assign sleep stages to all epochs of a whole night sleep recording.

Subject number	Scoring run-time (seconds)
1	32.63
2	29.87
3	35.67
4	32.21
5	32.25
6	33.11
7	31.71
8	31.05
9	31.93
10	33.57
11	31.90
12	30.62
13	31.88
14	31.17
15	38.67
16	32.87
17	30.68
18	32.83
19	33.41

6.2 Bibliography

1. Zammit, G. K., Weiner, J., Damato, N., Sillup, G. P., & McMillan, C. A. (1999). Quality of life in people with insomnia. *Sleep*, *22 Suppl 2*, S379–385.
2. Institute of Medicine. (2006). *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. Washington, DC: The National Academies Press. Retrieved April 5,2016 from http://www.nap.edu/catalog/11617/sleep-disorders-and-sleep-deprivation-an-unmet-public-health-problem
3. Natural Patterns of Sleep | A resource from the Division of Sleep Medicine at Harvard Medical School. (2007, December 18). *The Division of Sleep Medicine at Harvard Medical School*. Retrieved April 12, 2016, from http://healthysleep.med.harvard.edu/healthy/science/what/sleep-patterns-rem-nrem
4. Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Marcus, C. L., & Vaughn, B. V. (2012). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.0*. Darien, Illinois: American Academy of Sleep Medicine.
5. Danker-Hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., Dorffner, G. (2009). Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research*, *18*(1), 74–84. doi:10.1111/j.1365-2869.2008.00700.x
6. Silber, M. H., Ancoli-Israel, S., Bonnet, M. H., Chokroverty, S., Grigg-Damberger, M. M., Hirshkowitz, M., Iber, C. (2007). The Visual Scoring of Sleep in Adults. *Journal of Clinical Sleep Medicine*, *3*(2).
7. Stanus, E., Lacroix, B., Kerkhofs, M., & Mendlewicz, J. (1987). Automated sleep scoring: a comparative reliability study of two algorithms. *Electroencephalography and Clinical Neurophysiology*, *66*(4), 448–456. doi:10.1016/0013-4694(87)90214-8

8. Principe, J. C., Chang, T. G., Gala, S. K., & Tome, A. P. (1989). Information processing models for automatic sleep scoring. In *Engineering in Medicine and Biology Society, 1989. Images of the Twenty-First Century., Proceedings of the Annual International Conference of the IEEE Engineering in* (pp. 1804–1805 vol.6). doi:10.1109/IEMBS.1989.96466

9. Virkkala, J., Hasan, J., Värri, A., Himanen, S.-L., & Müller, K. (2007). Automatic sleep stage classification using two-channel electro-oculography. *Journal of Neuroscience Methods*, *166*(1), 109–115. doi:10.1016/j.jneumeth.2007.06.016

10. Liang, S. F., Kuo, C. E., Shaw, F. Z., Chen, Y. H., Hsu, C. H., & Chen, J. Y. (2015). Combination of expert knowledge and a genetic fuzzy inference system for automatic sleep staging. *IEEE Transactions on Biomedical Engineering*, doi:10.1109/TBME.2015.2510365

11. Gudmundsson, S., Runarsson, T. P., & Sigurdsson, S. (2005). Automatic Sleep Staging using Support Vector Machines with Posterior Probability Estimates. In *International Conference on Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce* (Vol. 2, pp. 366–372). doi:10.1109/CIMCA.2005.1631496

12. Oropesa, E., Cycon, H. L., & Jobert, M. (1999). Sleep Stage Classification using Wavelet Transform and Neural Network. *International Computer Science Institute*.

13. Doroshenkov, L. G., Konyshev, V. A., & Selishchev, S. V. (2007). Classification of human sleep stages based on EEG processing using hidden Markov models. *Biomedical Engineering*, *41*(1), 25–28. doi:10.1007/s10527-007-0006-5

14. Zoubek, L., Charbonnier, S., Lesecq, S., Buguet, A., & Chapotot, F. (2007). Feature selection for sleep/wake stages classification using data driven methods. *Biomedical Signal Processing and Control*, *2*(3), 171–179. doi:10.1016/j.bspc.2007.05.005

15. Sinha, R. K. (2008). Artificial Neural Network and Wavelet Based Automated Detection of Sleep Spindles, REM Sleep and Wake States. *Journal of Medical Systems*, *32*(4), 291–299. doi:10.1007/s10916-008-9134-z

16. Ebrahimi, F., Mikaeili, M., Estrada, E., & Nazeran, H. (2008). Automatic sleep stage classification based on EEG signals by using neural networks and wavelet packet coefficients. In *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008. EMBS 2008* (pp. 1151–1154). doi:10.1109/IEMBS.2008.4649365

17. Fraiwan, L. A., Khaswaneh, N. Y., & Lweesy, K. Y. (2009). Automatic Sleep Stage Scoring with Wavelet Packets Based on Single EEG Recording. *International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, *3*(6). Retrieved from http://waset.org/publications/9194/automatic-sleep-stage-scoring-with-wavelet-packets-based-on-single-eeg-recording

18. Güneş, S., Polat, K., & Yosunkaya, Ş. (2010). Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. *Expert Systems with Applications*, *37*(12), 7922–7928. doi:10.1016/j.eswa.2010.04.043

19. Jo, H. G., Park, J. Y., Lee, C. K., An, S. K., & Yoo, S. K. (2010). Genetic fuzzy classifier for sleep stage identification. *Computers in Biology and Medicine*, *40*(7), 629–634. doi:10.1016/j.compbiomed.2010.04.007

20. Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., & Dickhaus, H. (2012). Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, *108*(1), 10–19. doi:10.1016/j.cmpb.2011.11.005

21. Hsu, Y.-L., Yang, Y.-T., Wang, J.-S., & Hsu, C.-Y. (2013). Automatic sleep stage recurrent neural classifier using energy features of EEG signals. *Neurocomputing*, *104*, 105–114. doi:10.1016/j.neucom.2012.11.003

22. Liang, S. F., Kuo, C. E., Hu, Y. H., Pan, Y. H., & Wang, Y. H. (2012). Automatic Stage Scoring of Single-Channel Sleep EEG by Using Multiscale Entropy and Autoregressive Models. *IEEE Transactions on Instrumentation and Measurement*, *61*(6), 1649–1657. doi:10.1109/TIM.2012.2187242

23. Şen, B., Peker, M., Çavuşoğlu, A., & Çelebi, F. V. (2014). A Comparative Study on Classification of Sleep Stage Based on EEG Signals Using Feature Selection and

Classification Algorithms. *Journal of Medical Systems*, *38*(3), 1–21. doi:10.1007/s10916-014-0018-0

24. Lajnef, T., Chaibi, S., Ruby, P., Aguera, P.E., Eichenlaub, J.B., Samet, M., Jerbi, K. (2015). Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. *Journal of Neuroscience Methods*, *250*, 94–105. doi:10.1016/j.jneumeth.2015.01.022

25. Malaekah, E., Abdullah, H., & Cvetkovic, D. (2016). Automatic Sleep Stage Detection Based on Electrooculography. F. Ibrahim, J. Usman, S. M. Mohktar, & Y. M. Ahmad (Eds.), In *International Conference for Innovation in Biomedical Engineering and Life Sciences : ICIBEL2015, 6-8 December 2015, Putrajaya, Malaysia* (pp. 193–197). Singapore: Springer Singapore.

26. Hassan, A. R., & Bhuiyan, M. I. H. (2015). Automatic sleep stage classification. In *2015 2nd International Conference on Electrical Information and Communication Technology (EICT)* (pp. 211–216). doi:10.1109/EICT.2015.7391948

27. Pedro Fonseca and Xi Long and Mustafa Radha and Reinder Haakma and Ronald M Aarts and Jérôme Rolink. (2015). Sleep stage classification with ECG and respiratory effort. *Physiological Measurement*, *36*(10), 2027.

28. Yang, F., & Xia, B. (2016). Single Electrooculogram Channel-Based Sleep Stage Classification. R. Wang & X. Pan (Eds.), In *Advances in Cognitive Neurodynamics (V): Proceedings of the Fifth International Conference on Cognitive Neurodynamics - 2015* (pp. 595–600). Singapore: Springer Singapore.

29. Yaghouby, F., & Sunderam, S. (2015). Quasi-supervised scoring of human sleep in polysomnograms using augmented input variables. *Computers in Biology and Medicine*, *59*, 54–63. doi:10.1016/j.compbiomed.2015.01.012

30. Figueroa Helland, V. C., Gapelyuk, A., Suhrbier, A., Riedl, M., Penzel, T., Kurths, J., & Wessel, N. (2010). Investigation of an Automatic Sleep Stage Classification by Means of Multiscorer Hypnogram. *Methods of Information in Medicine*, *49*(5), 467–472. doi:10.3414/ME09-02-0052

31. Huang, C. S., Lin, C. L., Ko, L. W., Liu, S. Y., Sua, T. P., & Lin, C. T. (2013). A hierarchical classification system for sleep stage scoring via forehead EEG signals. In *2013 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)* (pp. 1–5). doi:10.1109/CCMB.2013.6609157

32. Imtiaz, S. A., & Rodriguez-Villegas, E. (2015). Automatic sleep staging using state machine-controlled decision trees. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 378–381). doi:10.1109/EMBC.2015.7318378

33. Tsinalis, O., Matthews, P. M., & Guo, Y. (2015). Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders. *Annals of Biomedical Engineering*, 1–11. doi:10.1007/s10439-015-1444-y

34. Tian, J. Y., & Liu, J. Q. (2005). Automated Sleep Staging by a Hybrid System Comprising Neural Network and Fuzzy Rule-based Reasoning. In *27th Annual International Conference of the Engineering in Medicine and Biology Society* (pp. 4115–4118). doi:10.1109/IEMBS.2005.1615368

35. Ma, H., Jackson, M., Yan, J., & Zhao, W. (2011). A Hybrid Classification Method using Artificial Neural Network Based Decision Tree for Automatic Sleep Scoring. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, *5*(7).

36. Park, H., Park, K., & Jeong, D.-U. (2000). Hybrid neural-network and rule-based expert system for automatic sleep stage scoring. In *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2000* (Vol. 2, pp. 1316–1319 vol.2). doi:10.1109/IEMBS.2000.897979

37. Bódizs, R., Sverteczki, M., & Mészáros, E. (2008). Wakefulness–sleep transition: Emerging electroencephalographic similarities with the rapid eye movement phase. *Brain Research Bulletin*, *76*(1–2), 85–89. doi:10.1016/j.brainresbull.2007.11.013

38. Guillaume, B., Charbonnier, S., Chapotot, F., Buguet, A., Bourdon, L., & Baconnier, P. (2005). Comparison Between Five Classifiers for Automatic Scoring of Human Sleep Recordings. *Studies in Computational Intelligence (SCI)*, *4*, 113–127.

39. Jobert, M., Tismer, C., Poiseau, E., & Schulz, H. (1994). Wavelets - A new tool in sleep biosignal analysis. *Journal of Sleep Research*, *3*, 223–232.

40. Robert, C., Guilpin, C., & Limoge, A. (1998). Review of neural network applications in sleep research. *Journal of Neuroscience Methods*, *79*, 187–193.

41. Buysse, D. J., Reynolds III, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research*, *28*(2), 193–213. doi:10.1016/0165-1781(89)90047-4

42. Estrada, E., Nazeran, H., Barragan, J., Burk, J. R., Lucas, E. A., & Behbehani, K. (2006). EOG and EMG: Two Important Switches in Automatic Sleep Stage Classification. In *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2006. EMBS '06* (pp. 2458–2461). doi:10.1109/IEMBS.2006.260075

43. Wendt, S. L., Christensen, J. A. E., Kempfner, J., Leonthin, H. L., Jennum, P., & Sorensen, H. B. D. (2012). Validation of a novel automatic sleep spindle detector with high performance during sleep in middle aged subjects. *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2012* (pp. 4250–4253). doi:10.1109/EMBC.2012.6346905

44. Rosenberg, R. S., & Van Hout, S. (2013). The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, *9*(1), 81–87. doi:10.5664/jcsm.2350

7 Curriculum Vitae

KRISTIN MARIA GUNNARSDOTTIR
kgunnar1@jhu.edu • (443)-301-7129

BIRTH DATE AND PLACE

October 29th 1990	Boston, Massachusetts

EDUCATION

Expected Spring 2016	**Master of Science in Biomedical Engineering** Johns Hopkins University, Baltimore, Maryland
Spring 2013	**Bachelor of Science in Biomedical Engineering** Reykjavik University, Reykjavik, Iceland

HONORS AND AWARDS

2013	**Certificate of appreciation for valuable contributions to Reykjavik University** Reykjavik University, Reykjavik, Iceland
2010-2013	**Dean's list, all semesters** Reykjavik University, Reykjavik, Iceland
2011	**Award for achieving the best results in physics over the whole country** The Physics Association of Iceland, Reykjavik, Iceland
2010	**Freshman's grant** Reykjavik University, Reykjavik, Iceland

WORK AND RESEARCH EXPERIENCE

August 2014 - Present — **Graduate Research Assistant**

Analyzing polysomnography data for automatic sleep stage detection and diagnosis of sleep disorders

Supervisor: Dr. Sridevi V. Sarma
Neuromedical Control Systems Lab
Johns Hopkins University, Baltimore, Maryland

May 2013 - May 2014 — **Research Technician**

Estimated cabin air quality in commercial aircrafts

Supervisor: Þorgeir Pálsson
Reykjavik University, Reykjavik, Iceland

January 2013 - April 2013 — **Engineering Intern**

Estimated signal quality of RIP breathing belts

Research and Development
Nox Medical, Reykjavik, Iceland

May 2012 - August 2012 — **Summer employee**

Projects related to mechanical prosthetic knees

Research & Development
Ossur hf., Reykjavik, Iceland

May 2011 — **Office Assistant**

Worked on the school's database of graduate students
School of Science and Engineering
Reykjavik University, Reykjavik, Iceland

TEACHING EXPERIENCE

August 2014 - Present — **Teaching Assistant**

Prepared students for lab work, graded homework assignments and lab reports

Johns Hopkins University, Baltimore, Maryland

January 2012 -	**Teaching Assistant**
December 2013	Conducted problem solving lectures, graded exams and homework assignments *Reykjavik University, Reykjavik, Iceland*
January 2012 -	**Instructor**
December 2013	Assisted first year students with homework related problems *Reykjavik University, Reykjavik, Iceland*

CERTIFICATES

2013	**Dale Carnegie Training** Dale Carnegie Training, Reykjavik, Iceland

PUBLICATIONS

Gunnarsdottir, K. M., Kang, Y. M., Kerr, M. S. D., Sarma, S. V., Ewen, J., Allen, R., Gamaldo, C., Salas, R. M. E. (2015). A look at the strength of micro and macro EEG analysis for distinguishing insomnia within an HIV cohort. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 6622–6625).

Kang, Y. M., **Gunnarsdottir, K. M.,** Kerr, M. S. D., Salas, R. M. E., Ewen, J., Allen, R., Gamaldo, C., Sarma, S. V. (2015). To Score or Not to Score? A look at the distinguishing power of micro EEG analysis on an annotated sample of PSG studies conducted in an HIV cohort. *In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 6626–6629).

PRESENTATIONS

ORAL PRESENTATIONS

Gunnarsdottir, K.M., Kang, Y.M., Salas, R.M.E, Gamaldo, C.E., Sarma, S.V. (June 2015). *Spatio- Temporal Dynamics of Sleep EEG in a Seropositive HIV Cohort.* Presented at CFAR Annual Meeting. Center for AIDS Research, Johns Hopkins University, Baltimore, Maryland.

POSTER PRESENTATIONS

Gunnarsdottir, K. M., Kang, Y. M., Kerr, M. S. D., Sarma, S. V., Ewen, J., Allen, R., Gamaldo, C., Salas, R. M. E. (August 2015). *A look at the strength of micro and macro EEG analysis for distinguishing insomnia within an HIV cohort*. Poster presented at the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy.

Kang, Y. M., **Gunnarsdottir, K. M.,** Kerr, M. S. D., Salas, R. M. E., Ewen, J., Allen, R., Gamaldo, C., Sarma, S. V. (August 2015). *To Score or Not to Score? A look at the distinguishing power of micro EEG analysis on an annotated sample of PSG studies conducted in an HIV cohort.* Poster presented at the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy.

Gunnarsdottir, K. M., Kang, Y. M., Kerr, M. S. D., Sarma, S. V., Ewen, J., Allen, R., Gamaldo, C., Salas, R. M. E. (June 2015). *A look at the strength of micro and macro EEG analysis for distinguishing insomnia within an HIV cohort*. Poster presented at Johns Hopkins University 1st Annual Sleep and Circadian Research Day, Baltimore, Maryland.

Kang, Y. M., **Gunnarsdottir, K. M.,** Kerr, M. S. D., Salas, R. M. E., Ewen, J., Allen, R., Gamaldo, C., Sarma, S. V. (June 2015). *To Score or Not to Score? A look at the distinguishing power of micro EEG analysis on an annotated sample of PSG studies conducted in an HIV cohort.* Poster presented at Johns Hopkins University 1st Annual Sleep and Circadian Research Day, Baltimore, Maryland.

Gunnarsdottir, K. M., Kang, Y. M., Kerr, M. S. D., Sarma, S. V., Ewen, J., Allen, R., Gamaldo, C., Salas, R. M. E. (May 2015). *A look at the strength of micro and macro EEG analysis for distinguishing insomnia within an HIV cohort*. Poster presented at the Seventh International Workshop Statistical Analysis of Neuronal Data (SAND7), Pittsburgh, Pennsylvania.